Trabajo de Control
Sistemas de Recuperación de Información
Fecha: 15 de julio de 2024
Temario #1

Nombre: _____ Grupo: _____

Seleccione en cada caso la(s) respuesta(s) correcta(s). Considere también no marcar ninguna opción.

1. Para transformar el contenido no estructurado de las páginas web en datos estructurados, el Web Scraping:
   a) Extrae información basada en patrones de HTML/CSS.
   b) Analiza los protocolos de red.
   c) Interpreta el código JavaScript en tiempo real.
   d) Convierte de forma automática imágenes a texto.

2. Implementar índices invertidos en un SRI asegura:
   a) Acelerar el proceso de búsqueda al permitir búsquedas directas por contenido en lugar de por título.
   b) Reducir la cantidad de espacio de almacenamiento necesario al comprimir los datos de los documentos.
   c) Facilitar la búsqueda sobre los datos que contienen términos específicos al mantener una lista de datos para cada término único.
   d) Incrementar la seguridad de los datos almacenados al dificultar el acceso directo a la información sin el índice correcto.

3. El propósito de la política de revisitado en los Web Crawlers es:
   a) Determinar la frecuencia con la que los crawlers deben volver a visitar una URL específica para mantener la información actualizada en el índice de búsqueda.
   b) Establecer reglas sobre el tiempo máximo que los crawlers pueden pasar en un sitio web durante cada visita para evitar sobrecargar los servidores.
   c) Limitar el acceso de los crawlers a ciertos servidores luego de visitar las páginas alojadas en estos.
   d) Definir la prioridad de rastreo de las URLs, determinando el orden en que los crawlers visitan y procesan cada página web.

4. El concepto de *rank sink* en el algoritmo de PageRank representa páginas web:
   a) Con distribuciones uniformes de la puntuación de PageRank.
   b) Que tienen una puntuación de PageRank más alta que otras debido a la manipulación de enlaces entrantes y salientes.
   c) Con una baja calidad de contenido y una cantidad insuficiente de enlaces salientes, lo que las hace menos relevantes en los resultados de búsqueda.
   d) Con un alto número de enlaces salientes que no reciben enlaces entrantes, lo que puede afectar negativamente su puntuación de PageRank.

5. ¿Cuál de las siguientes opciones describe mejor la diferencia clave entre Web Crawling y Web Scraping?
   a) Web Crawling se centra en la exploración y recopilación de enlaces de múltiples sitios web, mientras que Web Scraping se enfoca en la extracción específica de datos de páginas web individuales.
   b) Web Crawling se realiza utilizando herramientas de automatización como Selenium WebDriver, mientras que Web Scraping se lleva a cabo mediante el análisis de HTML y CSS.
   c) Web Scraping implica el análisis de la estructura y el contenido de las páginas web para extraer datos, mientras que Web Crawling se refiere a la descarga y almacenamiento de páginas web completas.
   d) Web Scraping es más eficaz para rastrear e indexar contenido web para motores de búsqueda, mientras que Web Crawling se utiliza principalmente para la extracción de datos en proyectos de investigación.

6. La transición de la Web 1.0 a la Web 2.0 se caracterizó principalmente por:
   a) El aumento en la velocidad de conexión a internet, que permitió una mejor calidad de las páginas web.
   b) La disminución de la importancia de los motores de búsqueda en la navegación web.
   c) El cambio de páginas web estáticas a dinámicas, permitiendo la interacción del usuario y la generación de contenido.
   d) La reducción en el uso de HTML y CSS en el desarrollo de sitios web.

7. Si una red cumple la propiedad de ser un grafo de mundo pequeño, entonces se conoce que:
   a) El número de componentes fuertemente conexas está relacionado con la cantidad de grafos K_n presentes.
   b) La longitud media del camino entre todo par de vértices es pequeña.

c) El grafo es un anillo regular de grado 5.

d) La red posee pocos vértices.

8. ¿Qué es un "Uniform Resource Locator (URL) Frontier" en el contexto de Web Crawling?

   a) Un protocolo que define cómo se deben formatear las URLs para el crawling.

   b) Una lista prioritaria de URLs que aún no han sido visitadas por el crawler.

   c) Una técnica para filtrar URLs irrelevantes y mejorar la eficiencia del crawling.

   d) Una base de datos que almacena URLs únicas identificadas como recursos en la web.

9. En la RI en el contexto de Big Data, se puede asegurar que:

   a) El almacenamiento distribuido centraliza todos los datos en un único servidor para facilitar su gestión y mantenimiento.

   b) El almacenamiento distribuido ofrece ventajas significativas en términos de escalabilidad y rendimiento en comparación con el almacenamiento centralizado.

   c) El almacenamiento distribuido reparte los datos en múltiples servidores para mejorar la disponibilidad y la redundancia del sistema.

   d) El almacenamiento distribuido es una técnica obsoleta.

10. En una biblioteca digital se necesita implementar un sistema que permita a los usuarios encontrar libros y artículos científicos de forma rápida y precisa. Los documentos están en diversos formatos, incluyendo PDF, EPUB y HTML. Se requiere seleccionar un algoritmo de indexación adecuado para el sistema, por lo que el programador designado para la implementación debe considerar:

    a) La velocidad de indexación y recuperación de datos.

    b) La capacidad para manejar documentos en diferentes formatos.

    c) La complejidad del algoritmo en términos de implementación y mantenimiento.

    d) La capacidad del algoritmo para procesar imágenes incrustadas o referenciadas en los ficheros.

11. En un sistema de control de tráfico urbano basado en reglas, ¿cuál de las siguientes reglas sería más efectiva para manejar situaciones de congestión en una intersección?

    a) Si hay pocos vehículos en la intersección, reducir el tiempo de los semáforos en verde.

    b) Si hay muchos vehículos en la intersección, aumentar el tiempo de los semáforos en verde.

    c) Si hay un vehículo de emergencia en la intersección, detener todos los demás vehículos.

    d) Si hay muchos peatones cruzando la intersección, reducir el tiempo de los semáforos en rojo.

12. Una buena práctica de SEO para mejorar el posicionamiento de un sitio web en los motores de búsqueda es:

    a) Obtener enlaces de otros sitios web relevantes y de calidad que apunten al sitio.

    b) Copiar contenido directamente de otros sitios web populares para aumentar la cantidad de páginas indexadas.

    c) Utilizar técnicas de encubrimiento para mostrar contenido diferente a lo indexado por los motores de búsqueda.

    d) Llenar el contenido con palabras clave irrelevantes para aumentar la densidad de palabras clave.

13. Dentro del análisis de redes, la centralidad de grado mide la importancia de un nodo basándose en:

    a) El grado del nodo.

    b) La cantidad de veces que aparece el nodo en el camino mínimo entre cualquier par de nodos.

    c) La cantidad de vecinos del nodo.

    d) El número de aristas que posee el nodo.

14. En la RI en el contexto de Big Data, se puede asegurar que:

    a) La indexación distribuida divide los datos en múltiples fragmentos que se almacenan en varios nodos para permitir búsquedas paralelas y mejorar la escalabilidad.

    b) La indexación distribuida no ofrece ventajas en términos de rendimiento y escalabilidad en comparación con la indexación centralizada.

    c) La indexación distribuida es una técnica obsoleta.

    d) La indexación distribuida almacena todos los datos en un solo servidor para facilitar su acceso y búsqueda.

15. En el algoritmo de Indexación Basada en Clasificación Bloqueada (BSBI), ¿cuál es el paso final para crear un índice invertido para la colección completa de los datos?

    a) Eliminar los términos duplicados de los índices.

    b) Indexar cada bloque de forma independiente.

    c) Fusionar los índices invertidos de cada bloque.

    d) Dividir la colección de datos en bloques de tamaño fijo.

16. Considerando las prácticas éticas y legales en el Web Scraping, se puede asegurar que:

    a) Web Scraping requiere considerar las políticas de `robots.txt` del sitio web objetivo.

b) Es importante revisar y respetar los términos de servicio del sitio web, así como las leyes aplicables de protección de datos y derechos de autor, antes de realizar Web Scraping.

c) Web Scraping sobre datos personales sin consentimiento es generalmente aceptado si los datos se utilizan con fines de investigación.

d) La extracción de datos mediante Web Scraping siempre es legal, independientemente de las leyes locales sobre derechos de autor y privacidad de datos definidos en los sitios web.

17. El conjunto frontera de URLs en un Web Crawler:

a) Delimita los sitios web que visitará en futuras iteraciones del proceso.

b) Es similar a un conjunto de URLs que esperan ser visitadas.

c) Almacena hipervínculos que pertenecen al mismo dominio del conjunto semilla de URLs con que inició el crawler.

d) Indica las secciones que no pueden ser visitadas de cada sitio web.

18. El algoritmo de PageRank puede describirse como un procedimiento utilizado para:

a) Determinar la velocidad de carga de una página web en un navegador.

b) Calcular la relevancia de una página web en función de la cantidad y calidad de los enlaces que apuntan hacia ella.

c) Prevenir el spam y el contenido no deseado en las páginas web.

d) Clasificar las páginas web en función de su edad y autoridad.

19. La centralidad de intermediación de un nodo indica:

a) La cantidad de nodos vecinos directos.

b) La resistencia del nodo a fallos.

c) El número total de conexiones entrantes y salientes.

d) La frecuencia con la que un nodo actúa como puente en el camino más corto entre otros dos nodos.

20. La diferencia entre la indexación por tokens y la indexación por conceptos puede definirse como:

a) La indexación por tokens asigna pesos a los términos basados en su importancia relativa, mientras que la indexación por conceptos utiliza un sistema de etiquetado para asociar términos con características generales.

b) La indexación por tokens divide los datos en términos individuales, mientras que la indexación por conceptos agrupa los datos en categorías definidas.

c) La indexación por tokens asigna un valor numérico a cada término de los datos, mientras que la indexación por conceptos utiliza algoritmos de encriptación para proteger la privacidad de los datos.

d) La indexación por tokens normaliza los datos reduciéndolos a su forma básica, mientras que la indexación por conceptos utiliza un método de ordenación para organizar los términos característicos de los datos.

Trabajo de Control
Sistemas de Recuperación de Información
Fecha: 15 de julio de 2024
Temario #2

Nombre: _____ Grupo: _____

Seleccione en cada caso la(s) respuesta(s) correcta(s). Considere también no marcar ninguna opción.

1. Se puede afirmar que:
   a) El término "Big Data" se refiere exclusivamente al volumen de datos que una organización maneja, sin tener en cuenta la velocidad, la variedad y la veracidad de los datos.
   b) Los SRI pueden beneficiarse de MapReduce para mejorar la RI relevante.
   c) MapReduce es un modelo de procesamiento distribuido utilizado para trabajar con grandes volúmenes de datos.
   d) Uno de los desafíos en el procesamiento de Big Data es la capacidad de gestionar y analizar datos provenientes de diversas fuentes y en diferentes formatos de manera eficiente.

2. El algoritmo de PageRank puede describirse como un procedimiento utilizado para:
   a) Determinar la velocidad de carga de una página web en un navegador.
   b) Calcular la relevancia de una página web en función de la cantidad y calidad de los enlaces que apuntan hacia ella.
   c) Prevenir el spam y el contenido no deseado en las páginas web.
   d) Clasificar las páginas web en función de su edad y autoridad.

3. El conjunto frontera de URLs en un Web Crawler:
   a) Delimita los sitios web que visitará en futuras iteraciones del proceso.
   b) Es similar a un conjunto de URLs que esperan ser visitadas.
   c) Almacena hipervínculos que pertenecen al mismo dominio del conjunto semilla de URLs con que inició el crawler.
   d) Indica las secciones que no pueden ser visitadas de cada sitio web.

4. Una buena práctica de SEO para mejorar el posicionamiento de un sitio web en los motores de búsqueda es:
   a) Obtener enlaces de otros sitios web relevantes y de calidad que apunten al sitio.
   b) Copiar contenido directamente de otros sitios web populares para aumentar la cantidad de páginas indexadas.
   c) Utilizar técnicas de encubrimiento para mostrar contenido diferente a lo indexado por los motores de búsqueda.
   d) Llenar el contenido con palabras clave irrelevantes para aumentar la densidad de palabras clave.

5. La integración de Hadoop y MapReduce en la RI trae como ventaja:
   a) La eliminación de la necesidad de sistemas de bases de datos.
   b) La posibilidad de la extracción de información relevante y la generación de resultados significativos de grandes conjuntos de datos.
   c) La garantía de la privacidad absoluta de los datos procesados.
   d) La reducción de los costos operativos a cero.

6. Al realizar la optimización de contenido para SEO debe considerarse:
   a) Seleccionar las palabras clave al azar.
   b) Utilizar etiquetas de título y meta descripciones únicas y relevantes para cada página.
   c) Incluir palabras clave de manera excesiva en el contenido para mejorar el posicionamiento.
   d) Crear contenido valioso y original que satisfaga las necesidades de los usuarios.

7. ¿Qué es un "Uniform Resource Locator (URL) Frontier" en el contexto de Web Crawling?
   a) Un protocolo que define cómo se deben formatear las URLs para el crawling.
   b) Una lista prioritaria de URLs que aún no han sido visitadas por el crawler.
   c) Una técnica para filtrar URLs irrelevantes y mejorar la eficiencia del crawling.
   d) Una base de datos que almacena URLs únicas identificadas como recursos en la web.

8. Analizar una red permite:
   a) Detectar posibles tendencias antes de que se conviertan en tendencia.

b) Obtener predicciones exactas de eventos futuros en mercados financieros.

c) Evaluar la calidad del contenido de un sitio web.

d) Encontrar nodos "sensibles" o críticos para la red.

9. Los Web Crawlers se enfrentan a desafíos constantes. Dentro de ellos se encuentran:

a) La dificultad para generar el contenido dinámico en tiempo real.

b) La modificación del código y la estructura del sitio web.

c) La incapacidad para interpretar correctamente el lenguaje de programación utilizado en el desarrollo de los sitios web.

d) La falta de acceso a la base de datos del servidor web para extraer información actualizada.

10. En un sistema de control de tráfico urbano basado en reglas, ¿cuál de las siguientes reglas sería más efectiva para manejar situaciones de congestión en una intersección?

a) Si hay pocos vehículos en la intersección, reducir el tiempo de los semáforos en verde.

b) Si hay muchos vehículos en la intersección, aumentar el tiempo de los semáforos en verde.

c) Si hay un vehículo de emergencia en la intersección, detener todos los demás vehículos.

d) Si hay muchos peatones cruzando la intersección, reducir el tiempo de los semáforos en rojo.

11. En el contexto de la RI en redes, se puede afirmar que:

a) La detección de comunidades no es relevante para la RI en redes, ya que se centra únicamente en la estructura de la red sin considerar el contenido.

b) La detección de comunidades en una red siempre produce resultados objetivos y consistentes independientemente del algoritmo utilizado.

c) La detección de comunidades ayuda a identificar grupos de nodos altamente conectados entre sí, lo que puede ser útil para comprender la estructura y el contenido de la red.

d) La detección de comunidades solo se aplica a redes pequeñas y simples, no a redes grandes y complejas.

12. En un grafo una comunidad es:

a) Un conjunto de nodos que no comparten ninguna similitud estructural o funcional entre sí.

b) Un conjunto de nodos aislados.

c) Un conjunto de nodos altamente conectados que forman un subgrafo completamente independiente del resto de la red.

d) Un conjunto de nodos que están más densamente interconectados entre sí que con los nodos fuera del conjunto.

13. En la RI en el contexto de Big Data, se puede asegurar que:

a) La indexación distribuida divide los datos en múltiples fragmentos que se almacenan en varios nodos para permitir búsquedas paralelas y mejorar la escalabilidad.

b) La indexación distribuida no ofrece ventajas en términos de rendimiento y escalabilidad en comparación con la indexación centralizada.

c) La indexación distribuida es una técnica obsoleta.

d) La indexación distribuida almacena todos los datos en un solo servidor para facilitar su acceso y búsqueda.

14. La premisa básica del algoritmo de PageRank para clasificar páginas web en los resultados de búsqueda es:

a) La antigüedad de la página web es el principal factor para determinar su clasificación.

b) El contenido y la relevancia de las palabras clave en la página web determinan su posición.

c) Los enlaces entrantes a una página web desde otras páginas contribuyen a su importancia y clasificación.

d) La cantidad de visitas que recibe una página web determina su clasificación.

15. No se considera como técnica para detectar comunidades en una red:

a) Analizar la mutualidad de los enlaces.

b) Usar el agrupamiento jerárquico.

c) Utilizar el algoritmo de K-Means.

d) Encontrar cliques de vértices de grado par.

16. La computación evolutiva:

a) No es aplicable en la RI debido a la complejidad de los algoritmos evolutivos.

b) Solo puede manejar conjuntos de datos pequeños y no es escalable a grandes volúmenes de datos.

c) Utiliza algoritmos para buscar soluciones óptimas en grandes espacios de búsqueda, lo que la hace adecuada para problemas de optimización en la RI.

d) Es útil solo para problemas de clasificación de documentos y no para otras tareas de RI en general.

17. Sobre WordNet y su aplicación en el procesamiento del lenguaje natural, se puede afirmar que:

a) El diseño de WordNet facilita su integración en aplicaciones multilingües de NLP, aunque su desarrollo original se centró en el inglés.

b) Aunque WordNet es una herramienta valiosa en el NLP, su estructura no incluye información sobre la frecuencia de uso de las palabras en el lenguaje natural.

c) Los synsets facilitan la identificación de relaciones semánticas entre palabras, como la hiperonimia y la meronimia, enriqueciendo tareas de NLP.

d) WordNet proporciona una base para la desambiguación semántica al agrupar palabras con significados similares en synsets.

18. ¿Cuál de las siguientes opciones describe mejor la diferencia clave entre Web Crawling y Web Scraping?

a) Web Crawling se centra en la exploración y recopilación de enlaces de múltiples sitios web, mientras que Web Scraping se enfoca en la extracción específica de datos de páginas web individuales.

b) Web Crawling se realiza utilizando herramientas de automatización como Selenium WebDriver, mientras que Web Scraping se lleva a cabo mediante el análisis de HTML y CSS.

c) Web Scraping implica el análisis de la estructura y el contenido de las páginas web para extraer datos, mientras que Web Crawling se refiere a la descarga y almacenamiento de páginas web completas.

d) Web Scraping es más eficaz para rastrear e indexar contenido web para motores de búsqueda, mientras que Web Crawling se utiliza principalmente para la extracción de datos en proyectos de investigación.

19. En una biblioteca digital se necesita implementar un sistema que permita a los usuarios encontrar libros y artículos científicos de forma rápida y precisa. Los documentos están en diversos formatos, incluyendo PDF, EPUB y HTML. Se requiere seleccionar un algoritmo de indexación adecuado para el sistema, por lo que el programador designado para la implementación debe considerar:

a) La velocidad de indexación y recuperación de datos.

b) La capacidad para manejar documentos en diferentes formatos.

c) La complejidad del algoritmo en términos de implementación y mantenimiento.

d) La capacidad del algoritmo para procesar imágenes incrustadas o referenciadas en los ficheros.

20. Al "relajar" el concepto de clique en la detección de comunidades se intenta solucionar:

a) La imposibilidad de conectar nodos distantes.

b) La necesidad de datos externos para analizar la red.

c) La uniformidad de los nodos en términos de grado.

d) El solapamiento y la complejidad computacional.

Trabajo de Control
Sistemas de Recuperación de Información
Fecha: 15 de julio de 2024
Temario #3

Nombre: _____ Grupo: _____

Seleccione en cada caso la(s) respuesta(s) correcta(s). Considere también no marcar ninguna opción.

1. Sobre el algoritmo de PageRank, visto en clase, se puede afirmar que:
   a) Evalúa la importancia de un sitio web en función de la calidad y cantidad de enlaces entrantes que recibe de otros sitios web.
   b) Solo tiene en cuenta el contenido en un sitio web para determinar su relevancia en los resultados de búsqueda.
   c) Asigna una puntuación alta a los sitios web que tienen un gran número de enlaces entrantes sin tener en cuenta la calidad de esos enlaces.
   d) Asigna una puntuación baja a los sitios web que contienen muchos enlaces salientes, ya que indica una falta de relevancia.

2. Se puede afirmar que:
   a) El término "Big Data" se refiere exclusivamente al volumen de datos que una organización maneja, sin tener en cuenta la velocidad, la variedad y la veracidad de los datos.
   b) Los SRI pueden beneficiarse de MapReduce para mejorar la RI relevante.
   c) MapReduce es un modelo de procesamiento distribuido utilizado para trabajar con grandes volúmenes de datos.
   d) Uno de los desafíos en el procesamiento de Big Data es la capacidad de gestionar y analizar datos provenientes de diversas fuentes y en diferentes formatos de manera eficiente.

3. ¿Qué algoritmos permiten obtener información de una red?
   a) Índices de centralidad.
   b) Detección de comunidades.
   c) Hypertext Induced Topic Selection (HITS).
   d) PageRank.

4. Para contribuir positivamente al posicionamiento orgánico de un sitio web en los motores de búsqueda se puede:
   a) Crear contenido relevante y de alta calidad que satisfaga las necesidades de información de los usuarios.
   b) Mejorar la velocidad de carga del sitio web y asegurar que sea *responsive* y fácil de usar en dispositivos móviles.
   c) Incluir una densidad alta de palabras clave para asegurar que el sitio web aparezca en tantas búsquedas como sea posible.
   d) Obtener enlaces entrantes de otros sitios web de alta autoridad y relevancia temática.

5. Considerando las prácticas éticas y legales en el Web Scraping, se puede asegurar que:
   a) Web Scraping requiere considerar las políticas de `robots.txt` del sitio web objetivo.
   b) Es importante revisar y respetar los términos de servicio del sitio web, así como las leyes aplicables de protección de datos y derechos de autor, antes de realizar Web Scraping.
   c) Web Scraping sobre datos personales sin consentimiento es generalmente aceptado si los datos se utilizan con fines de investigación.
   d) La extracción de datos mediante Web Scraping siempre es legal, independientemente de las leyes locales sobre derechos de autor y privacidad de datos definidos en los sitios web.

6. Si se tiene el conjunto de páginas interconectadas $\{A \rightarrow B, C; B \rightarrow C; C \rightarrow A\}$, entonces la página con valor más alto de PageRank es:
   a) A y B.
   b) A.
   c) C.
   d) B.

7. Al realizar la optimización de contenido para SEO debe considerarse:
   a) Seleccionar las palabras clave al azar.
   b) Utilizar etiquetas de título y meta descripciones únicas y relevantes para cada página.
   c) Incluir palabras clave de manera excesiva en el contenido para mejorar el posicionamiento.
   d) Crear contenido valioso y original que satisfaga las necesidades de los usuarios.

8. En el modelo de representación del conocimiento basado en herencia se puede asegurar que:

a) La cantidad de padres que puede tener un nodo no es mayor que 4.

b) Las conclusiones pueden ser canceladas si el grafo es ambiguo.

c) Solo se usa en entornos referentes a la biología.

d) El razonamiento deducido está respaldado por al menos un camino dentro del grafo.

9. Se puede afirmar que:

a) No es posible identificar subgrupos dentro de una red utilizando análisis de redes.

b) La cantidad de conexiones de un nodo siempre indica su influencia en la red.

c) El tamaño de una red es siempre indicativo de su efectividad en la transmisión de información.

d) Todas las relaciones en una red tienen la misma importancia para el análisis.

10. Dentro del ecosistema de Hadoop, el HDFS se caracteriza por:

a) El modelo de acceso y de escritura de datos en tiempo real.

b) La tolerancia a fallos mediante la replicación de datos.

c) El almacenamiento exclusivo para archivos de texto.

d) La capacidad ilimitada de almacenamiento.

11. No se considera como técnica para detectar comunidades en una red:

a) Analizar la mutualidad de los enlaces.

b) Usar el agrupamiento jerárquico.

c) Utilizar el algoritmo de K-Means.

d) Encontrar cliques de vértices de grado par.

12. ¿Qué estrategia utilizan los Web Crawlers para asegurar un rastreo eficiente y respetuoso de los recursos de los sitios web?

a) Ignorar completamente el archivo `robots.txt` de los sitios web.

b) Visitar y rastrear todos los enlaces de una página web simultáneamente.

c) Extraer únicamente contenido multimedia para reducir la carga en los servidores web.

d) Seguir las directrices del archivo `robots.txt` y aplicar un retraso entre las solicitudes.

13. En una red de transporte donde cada nodo es una parada de autobús y las aristas representan si existe un carro que pasa por ambos sitios, ¿qué puede mejorar el sistema de transporte?

a) No considerar la centralidad de intermediación de las estaciones de transporte público al planificar rutas y horarios, ya que no tiene impacto en las conexiones entre las paradas.

b) Utilizar el grafo inducido de los nodos con mayor valor en la centralidad de grado para aplicar la centralidad de intermediación con el propósito de reforzar las paradas con mayor tráfico.

c) Utilizar la centralidad de intermediación para identificar las paradas de transferencia clave y establecer nuevas conexiones entre diferentes líneas de transporte público.

d) No implementar sistemas de información en tiempo real para los usuarios, ya que pueden aumentar la carga de los trabajadores.

14. ¿Qué es un "Uniform Resource Locator (URL) Frontier" en el contexto de Web Crawling?

a) Un protocolo que define cómo se deben formatear las URLs para el crawling.

b) Una lista prioritaria de URLs que aún no han sido visitadas por el crawler.

c) Una técnica para filtrar URLs irrelevantes y mejorar la eficiencia del crawling.

d) Una base de datos que almacena URLs únicas identificadas como recursos en la web.

15. Un investigador necesita recopilar datos de múltiples sitios web para un estudio académico, pero se enfrenta a varios desafíos al realizar el proceso de Web Scraping de manera ética y legal. ¿Cuál de las siguientes opciones describe mejor uno de los desafíos asociados al proceso de Web Scraping?

a) La necesidad de comprender la estructura del sitio web y su código HTML para extraer los datos correctamente.

b) La disponibilidad limitada de datos en línea que se pueden extraer utilizando técnicas de Web Scraping.

c) La dificultad para encontrar herramientas de Web Scraping gratuitas y fiables.

d) La necesidad de estar montado sobre un scrawler que cumpla con todas las políticas.

16. En la RI en el contexto de Big Data, se puede asegurar que:

a) La indexación distribuida divide los datos en múltiples fragmentos que se almacenan en varios nodos para permitir búsquedas paralelas y mejorar la escalabilidad.

b) La indexación distribuida no ofrece ventajas en términos de rendimiento y escalabilidad en comparación con la indexación centralizada.

c) La indexación distribuida es una técnica obsoleta.

d) La indexación distribuida almacena todos los datos en un solo servidor para facilitar su acceso y búsqueda.

17. En el algoritmo de Indexación Basada en Clasificación Bloqueada (BSBI), ¿cuál es el paso final para crear un índice invertido para la colección completa de los datos?
    a) Eliminar los términos duplicados de los índices.
    b) Indexar cada bloque de forma independiente.
    c) Fusionar los índices invertidos de cada bloque.
    d) Dividir la colección de datos en bloques de tamaño fijo.

18. El concepto de *rank sink* en el algoritmo de PageRank representa páginas web:
    a) Con distribuciones uniformes de la puntuación de PageRank.
    b) Que tienen una puntuación de PageRank más alta que otras debido a la manipulación de enlaces entrantes y salientes.
    c) Con una baja calidad de contenido y una cantidad insuficiente de enlaces salientes, lo que las hace menos relevantes en los resultados de búsqueda.
    d) Con un alto número de enlaces salientes que no reciben enlaces entrantes, lo que puede afectar negativamente su puntuación de PageRank.

19. La Web 2.0 se caracteriza por:
    a) La existencia de sitios web dinámicos e interactivos que permiten a los usuarios participar, comentar e interactuar tanto con los creadores de contenido como con otros usuarios.
    b) El uso de distintas tecnologías para crear experiencias web más interactivas y con mayor capacidad de respuesta.
    c) Las actualizaciones del contenido de los sitios web se efectúa de forma manual dentro del código HTML
    d) Las plataformas ofrecen experiencias a la medida, permitiendo a los usuarios personalizar sus perfiles, recibir recomendaciones ajustadas al contenido y participar en filtrado colaborativo.

20. El propósito de la política de revisitado en los Web Crawlers es:
    a) Determinar la frecuencia con la que los crawlers deben volver a visitar una URL específica para mantener la información actualizada en el índice de búsqueda.
    b) Establecer reglas sobre el tiempo máximo que los crawlers pueden pasar en un sitio web durante cada visita para evitar sobrecargar los servidores.
    c) Limitar el acceso de los crawlers a ciertos servidores luego de visitar las páginas alojadas en estos.
    d) Definir la prioridad de rastreo de las URLs, determinando el orden en que los crawlers visitan y procesan cada página web.

Trabajo de Control
Sistemas de Recuperación de Información
Fecha: 15 de julio de 2024
Temario #4

Nombre: _____ Grupo: _____

Seleccione en cada caso la(s) respuesta(s) correcta(s). Considere también no marcar ninguna opción.

1. En la RI en el contexto de Big Data, se puede asegurar que:
   a) El almacenamiento distribuido centraliza todos los datos en un único servidor para facilitar su gestión y mantenimiento.
   b) El almacenamiento distribuido ofrece ventajas significativas en términos de escalabilidad y rendimiento en comparación con el almacenamiento centralizado.
   c) El almacenamiento distribuido reparte los datos en múltiples servidores para mejorar la disponibilidad y la redundancia del sistema.
   d) El almacenamiento distribuido es una técnica obsoleta.

2. La afirmación que mejor refleja el principio subyacente de PageRank, considerando su importancia en la clasificación de los sitios web, es:
   a) PageRank valora más la cantidad de enlaces entrantes a una página web, independientemente de la calidad o relevancia de estos enlaces.
   b) La efectividad de PageRank se basa exclusivamente en el análisis de las palabras clave contenidas en los enlaces entrantes, sin considerar la estructura de enlace de la Web.
   c) El algoritmo de PageRank considera tanto la cantidad como la calidad de los enlaces entrantes, asignando mayor valor a los enlaces provenientes de sitios web considerados como "importantes".
   d) PageRank opera bajo el supuesto de que los enlaces entrantes y salientes tienen el mismo impacto en la valoración de la relevancia de una página web.

3. Los Web Crawlers se enfrentan a desafíos constantes. Dentro de ellos se encuentran:
   a) La dificultad para generar el contenido dinámico en tiempo real.
   b) La modificación del código y la estructura del sitio web.
   c) La incapacidad para interpretar correctamente el lenguaje de programación utilizado en el desarrollo de los sitios web.
   d) La falta de acceso a la base de datos del servidor web para extraer información actualizada.

4. El posicionamiento de un sitio web en los motores de búsqueda puede ser afectado por:
   a) Utilizar técnicas de encubrimiento para mostrar contenido diferente a lo indexado por los motores de búsqueda.
   b) Contenido sin valor en el sitio web.
   c) Mantener una estructura de URL clara y coherente.
   d) Obtener enlaces de sitios web irrelevantes y de baja calidad.

5. El concepto de *rank sink* en el algoritmo de PageRank representa páginas web:
   a) Con distribuciones uniformes de la puntuación de PageRank.
   b) Que tienen una puntuación de PageRank más alta que otras debido a la manipulación de enlaces entrantes y salientes.
   c) Con una baja calidad de contenido y una cantidad insuficiente de enlaces salientes, lo que las hace menos relevantes en los resultados de búsqueda.
   d) Con un alto número de enlaces salientes que no reciben enlaces entrantes, lo que puede afectar negativamente su puntuación de PageRank.

6. En un sistema donde el conocimiento está definido a partir de reglas se puede asegurar que:
   a) El nivel de especificidad de las reglas está limitado.
   b) La representación de conocimiento está basada en la lógica proposicional.
   c) El razonamiento se activa en cadena hacia delante.
   d) El orden en que se definen las reglas no altera el razonamiento del sistema.

7. El algoritmo de PageRank puede describirse como un procedimiento utilizado para:
   a) Determinar la velocidad de carga de una página web en un navegador.
   b) Calcular la relevancia de una página web en función de la cantidad y calidad de los enlaces que apuntan hacia ella.
   c) Prevenir el spam y el contenido no deseado en las páginas web.
   d) Clasificar las páginas web en función de su edad y autoridad.

8. La Web 2.0 se caracteriza por:

a) La existencia de sitios web dinámicos e interactivos que permiten a los usuarios participar, comentar e interactuar tanto con los creadores de contenido como con otros usuarios.

b) El uso de distintas tecnologías para crear experiencias web más interactivas y con mayor capacidad de respuesta.

c) Las actualizaciones del contenido de los sitios web se efectúa de forma manual dentro del código HTML

d) Las plataformas ofrecen experiencias a la medida, permitiendo a los usuarios personalizar sus perfiles, recibir recomendaciones ajustadas al contenido y participar en filtrado colaborativo.

9. Para contribuir positivamente al posicionamiento orgánico de un sitio web en los motores de búsqueda se puede:

a) Crear contenido relevante y de alta calidad que satisfaga las necesidades de información de los usuarios.

b) Mejorar la velocidad de carga del sitio web y asegurar que sea *responsive* y fácil de usar en dispositivos móviles.

c) Incluir una densidad alta de palabras clave para asegurar que el sitio web aparezca en tantas búsquedas como sea posible.

d) Obtener enlaces entrantes de otros sitios web de alta autoridad y relevancia temática.

10. Al "relajar" el concepto de clique en la detección de comunidades se intenta solucionar:

a) La imposibilidad de conectar nodos distantes.

b) La necesidad de datos externos para analizar la red.

c) La uniformidad de los nodos en términos de grado.

d) El solapamiento y la complejidad computacional.

11. En el contexto de la RI en redes sociales se puede afirmar que:

a) El análisis de redes sociales se centra exclusivamente en la cantidad de seguidores que tiene un usuario en particular para determinar su influencia en la red.

b) El análisis de centralidad de intermediación se utiliza para identificar usuarios que son importantes en una red social debido a su posición como "puentes" entre diferentes grupos de usuarios.

c) El análisis de sentimientos se utiliza para determinar la popularidad de una publicación en redes sociales sin tener en cuenta la opinión de los usuarios.

d) El análisis de redes sociales no es útil para comprender la difusión de información en una red social específica, para ello se utiliza la medida de centralidad de vector propio.

12. La Web 3 se conoce como:

a) Internet de las cosas.

b) Web de solo lectura.

c) Web semántica.

d) Web de escritura-lectura.

13. La integración de Hadoop y MapReduce en la RI trae como ventaja:

a) La eliminación de la necesidad de sistemas de bases de datos.

b) La posibilidad de la extracción de información relevante y la generación de resultados significativos de grandes conjuntos de datos.

c) La garantía de la privacidad absoluta de los datos procesados.

d) La reducción de los costos operativos a cero.

14. ¿Qué estrategia utilizan los Web Crawlers para asegurar un rastreo eficiente y respetuoso de los recursos de los sitios web?

a) Ignorar completamente el archivo `robots.txt` de los sitios web.

b) Visitar y rastrear todos los enlaces de una página web simultáneamente.

c) Extraer únicamente contenido multimedia para reducir la carga en los servidores web.

d) Seguir las directrices del archivo `robots.txt` y aplicar un retraso entre las solicitudes.

15. En el modelo de representación del conocimiento basado en herencia se puede asegurar que:

a) La cantidad de padres que puede tener un nodo no es mayor que 4.

b) Las conclusiones pueden ser canceladas si el grafo es ambiguo.

c) Solo se usa en entornos referentes a la biología.

d) El razonamiento deducido está respaldado por al menos un camino dentro del grafo.

16. Sobre WordNet y su aplicación en el procesamiento del lenguaje natural, se puede afirmar que:

a) El diseño de WordNet facilita su integración en aplicaciones multilingües de NLP, aunque su desarrollo original se centró en el inglés.

b) Aunque WordNet es una herramienta valiosa en el NLP, su estructura no incluye información sobre la frecuencia de uso de las palabras en el lenguaje natural.

c) Los synsets facilitan la identificación de relaciones semánticas entre palabras, como la hiperonimia y la meronimia, enriqueciendo tareas de NLP.

d) WordNet proporciona una base para la desambiguación semántica al agrupar palabras con significados similares en synsets.

17. El conjunto frontera de URLs en un Web Crawler:

a) Delimita los sitios web que visitará en futuras iteraciones del proceso.

b) Es similar a un conjunto de URLs que esperan ser visitadas.

c) Almacena hipervínculos que pertenecen al mismo dominio del conjunto semilla de URLs con que inició el crawler.

d) Indica las secciones que no pueden ser visitadas de cada sitio web.

18. Los algoritmos para detectar comunidades en una red intentan:

a) Buscar conjuntos donde cada nodo de un mismo conjunto tenga características similares al resto de los nodos del conjunto.

b) Seleccionar aleatoriamente nodos de alto grado y sus vecinos.

c) Buscar subgrafos tal que no incluyan nodos cuya ausencia desconecte al subgrafo.

d) Encontrar grupos donde los nodos pertenecientes a los mismos grupos son cercanos bajo cierta métrica y lejanos con respecto a los nodos de otros grupos.

19. Con respecto a WordNet y su contribución a las ontologías y la representación del conocimiento en los sistemas de RI, puede afirmarse que:

a) En WordNet, una palabra está asociada a un conjunto de sinónimos (synsets), siendo estas palabras intercambiables en un contexto.

b) La integración de WordNet en sistemas de RI permite la expansión de consultas y la mejora de la precisión de los resultados al entender mejor el significado de los términos de búsqueda a través de su contexto semántico y las relaciones entre palabras.

c) WordNet diferencia claramente entre los significados de palabras según su uso en diferentes contextos, lo que permite aplicaciones avanzadas en desambiguación semántica más allá de los sistemas de recomendación y RI.

d) Una limitación de WordNet en la representación del conocimiento es su enfoque en el idioma inglés, lo que plantea desafíos en la aplicación global y la interoperabilidad con sistemas de información multilingües.

20. En un sistema de control de tráfico urbano basado en reglas, ¿cuál de las siguientes reglas sería más efectiva para manejar situaciones de congestión en una intersección?

a) Si hay pocos vehículos en la intersección, reducir el tiempo de los semáforos en verde.

b) Si hay muchos vehículos en la intersección, aumentar el tiempo de los semáforos en verde.

c) Si hay un vehículo de emergencia en la intersección, detener todos los demás vehículos.

d) Si hay muchos peatones cruzando la intersección, reducir el tiempo de los semáforos en rojo.

Trabajo de Control
Sistemas de Recuperación de Información
Fecha: 15 de julio de 2024
Temario #5

Nombre: _____ Grupo: _____

Seleccione en cada caso la(s) respuesta(s) correcta(s). Considere también no marcar ninguna opción.

1. La diferencia entre la indexación por tokens y la indexación por conceptos puede definirse como:
   a) La indexación por tokens asigna pesos a los términos basados en su importancia relativa, mientras que la indexación por conceptos utiliza un sistema de etiquetado para asociar términos con características generales.
   b) La indexación por tokens divide los datos en términos individuales, mientras que la indexación por conceptos agrupa los datos en categorías definidas.
   c) La indexación por tokens asigna un valor numérico a cada término de los datos, mientras que la indexación por conceptos utiliza algoritmos de encriptación para proteger la privacidad de los datos.
   d) La indexación por tokens normaliza los datos reduciéndolos a su forma básica, mientras que la indexación por conceptos utiliza un método de ordenación para organizar los términos característicos de los datos.

2. La afirmación que mejor refleja el principio subyacente de PageRank, considerando su importancia en la clasificación de los sitios web, es:
   a) PageRank valora más la cantidad de enlaces entrantes a una página web, independientemente de la calidad o relevancia de estos enlaces.
   b) La efectividad de PageRank se basa exclusivamente en el análisis de las palabras clave contenidas en los enlaces entrantes, sin considerar la estructura de enlace de la Web.
   c) El algoritmo de PageRank considera tanto la cantidad como la calidad de los enlaces entrantes, asignando mayor valor a los enlaces provenientes de sitios web considerados como "importantes".
   d) PageRank opera bajo el supuesto de que los enlaces entrantes y salientes tienen el mismo impacto en la valoración de la relevancia de una página web.

3. Al realizar la optimización de contenido para SEO debe considerarse:
   a) Seleccionar las palabras clave al azar.
   b) Utilizar etiquetas de título y meta descripciones únicas y relevantes para cada página.
   c) Incluir palabras clave de manera excesiva en el contenido para mejorar el posicionamiento.
   d) Crear contenido valioso y original que satisfaga las necesidades de los usuarios.

4. Sobre el algoritmo de PageRank, visto en clase, se puede afirmar que:
   a) Evalúa la importancia de un sitio web en función de la calidad y cantidad de enlaces entrantes que recibe de otros sitios web.
   b) Solo tiene en cuenta el contenido en un sitio web para determinar su relevancia en los resultados de búsqueda.
   c) Asigna una puntuación alta a los sitios web que tienen un gran número de enlaces entrantes sin tener en cuenta la calidad de esos enlaces.
   d) Asigna una puntuación baja a los sitios web que contienen muchos enlaces salientes, ya que indica una falta de relevancia.

5. Para la RI, el análisis de las redes puede:
   a) Ayudar a identificar grupos de interés.
   b) Indicar la importancia de una entidad en la transmisión de la información.
   c) Ayudar a comprender la conectividad y la accesibilidad entre las entidades.
   d) Revelar patrones de influencia dentro de una comunidad.

6. El concepto de *rank sink* en el algoritmo de PageRank representa páginas web:
   a) Con distribuciones uniformes de la puntuación de PageRank.
   b) Que tienen una puntuación de PageRank más alta que otras debido a la manipulación de enlaces entrantes y salientes.
   c) Con una baja calidad de contenido y una cantidad insuficiente de enlaces salientes, lo que las hace menos relevantes en los resultados de búsqueda.
   d) Con un alto número de enlaces salientes que no reciben enlaces entrantes, lo que puede afectar negativamente su puntuación de PageRank.

7. En un sistema de control de tráfico urbano basado en reglas, ¿cuál de las siguientes reglas sería más efectiva para manejar situaciones de congestión en una intersección?

a) Si hay pocos vehículos en la intersección, reducir el tiempo de los semáforos en verde.

b) Si hay muchos vehículos en la intersección, aumentar el tiempo de los semáforos en verde.

c) Si hay un vehículo de emergencia en la intersección, detener todos los demás vehículos.

d) Si hay muchos peatones cruzando la intersección, reducir el tiempo de los semáforos en rojo.

8. En un sistema donde el conocimiento está definido a partir de reglas se puede asegurar que:

a) El nivel de especificidad de las reglas está limitado.

b) La representación de conocimiento está basada en la lógica proposicional.

c) El razonamiento se activa en cadena hacia delante.

d) El orden en que se definen las reglas no altera el razonamiento del sistema.

9. ¿Qué estrategia utilizan los Web Crawlers para asegurar un rastreo eficiente y respetuoso de los recursos de los sitios web?

a) Ignorar completamente el archivo `robots.txt` de los sitios web.

b) Visitar y rastrear todos los enlaces de una página web simultáneamente.

c) Extraer únicamente contenido multimedia para reducir la carga en los servidores web.

d) Seguir las directrices del archivo `robots.txt` y aplicar un retraso entre las solicitudes.

10. La Web 1.0 se caracteriza por:

a) Las actualizaciones del contenido de los sitios web se efectúa de forma manual dentro del código HTML

b) Los sitios web se centran en brindar información en lugar de facilitar la colaboración o participación de los usuarios.

c) Uso de comunidades virtuales para popularizar los sitios web de noticias.

d) Los propietarios de los sitios web proporcionan contenido de forma periódica.

11. Se tiene un grafo donde cada nodo es un personaje de cierto libro de cuentos y la existencia de las aristas está definida si dos personajes aparecen en el mismo cuento. Se puede asegurar que:

a) La centralidad de intermediación identifica a los personajes que actúan como conectores entre personajes de cuentos distintos.

b) La centralidad de grado es útil para identificar los personajes con más conexiones dentro de la red, lo que puede indicar su importancia en el libro.

c) La centralidad de cercanía ofrece una relación entre la cantidad de vecinos de un nodo con respecto a la longitud máxima de un camino dentro del grafo partiendo del nodo en cuestión.

d) La centralidad de cercanía indica el grado de conexión de cada personaje con el resto de los personajes de los cuentos del libro.

12. Se puede afirmar que:

a) El término "Big Data" se refiere exclusivamente al volumen de datos que una organización maneja, sin tener en cuenta la velocidad, la variedad y la veracidad de los datos.

b) Los SRI pueden beneficiarse de MapReduce para mejorar la RI relevante.

c) MapReduce es un modelo de procesamiento distribuido utilizado para trabajar con grandes volúmenes de datos.

d) Uno de los desafíos en el procesamiento de Big Data es la capacidad de gestionar y analizar datos provenientes de diversas fuentes y en diferentes formatos de manera eficiente.

13. El uso de Hadoop y MapReduce en el contexto de la RI tiene como objetivo:

a) Automatizar el mantenimiento de sistemas de bases de datos.

b) Facilitar el análisis en tiempo real de datos de redes sociales.

c) Mejorar la eficiencia energética en centros de datos.

d) Procesar y analizar grandes conjuntos de datos para la RI.

14. Analizar una red permite:

a) Detectar posibles tendencias antes de que se conviertan en tendencia.

b) Obtener predicciones exactas de eventos futuros en mercados financieros.

c) Evaluar la calidad del contenido de un sitio web.

d) Encontrar nodos "sensibles" o críticos para la red.

15. Dentro del ecosistema de Hadoop, el HDFS se caracteriza por:

a) El modelo de acceso y de escritura de datos en tiempo real.

b) La tolerancia a fallos mediante la replicación de datos.

c) El almacenamiento exclusivo para archivos de texto.

d) La capacidad ilimitada de almacenamiento.

16. El propósito de la política de revisitado en los Web Crawlers es:

a) Determinar la frecuencia con la que los crawlers deben volver a visitar una URL específica para mantener la información actualizada en el índice de búsqueda.

b) Establecer reglas sobre el tiempo máximo que los crawlers pueden pasar en un sitio web durante cada visita para evitar sobrecargar los servidores.

c) Limitar el acceso de los crawlers a ciertos servidores luego de visitar las páginas alojadas en estos.

d) Definir la prioridad de rastreo de las URLs, determinando el orden en que los crawlers visitan y procesan cada página web.

17. En una biblioteca digital se necesita implementar un sistema que permita a los usuarios encontrar libros y artículos científicos de forma rápida y precisa. Los documentos están en diversos formatos, incluyendo PDF, EPUB y HTML. Se requiere seleccionar un algoritmo de indexación adecuado para el sistema, por lo que el programador designado para la implementación debe considerar:

a) La velocidad de indexación y recuperación de datos.

b) La capacidad para manejar documentos en diferentes formatos.

c) La complejidad del algoritmo en términos de implementación y mantenimiento.

d) La capacidad del algoritmo para procesar imágenes incrustadas o referenciadas en los ficheros.

18. La computación evolutiva:

a) No es aplicable en la RI debido a la complejidad de los algoritmos evolutivos.

b) Solo puede manejar conjuntos de datos pequeños y no es escalable a grandes volúmenes de datos.

c) Utiliza algoritmos para buscar soluciones óptimas en grandes espacios de búsqueda, lo que la hace adecuada para problemas de optimización en la RI.

d) Es útil solo para problemas de clasificación de documentos y no para otras tareas de RI en general.

19. Si una red cumple la propiedad de ser un grafo de mundo pequeño, entonces se conoce que:

a) El número de componentes fuertemente conexas está relacionado con la cantidad de grafos K_n presentes.

b) La longitud media del camino entre todo par de vértices es pequeña.

c) El grafo es un anillo regular de grado 5.

d) La red posee pocos vértices.

20. En un SRI la indexación:

a) Consiste en asociar un identificador único a cada dato almacenado en el sistema.

b) Mejora la experiencia del usuario.

c) Permite la organización y la categorización de la información.

d) Facilita la RI relevante.

Trabajo de Control
Sistemas de Recuperación de Información
Fecha: 15 de julio de 2024
Temario #6

Nombre: _____ Grupo: _____

Seleccione en cada caso la(s) respuesta(s) correcta(s). Considere también no marcar ninguna opción.

1. ¿Qué estrategia utilizan los Web Crawlers para asegurar un rastreo eficiente y respetuoso de los recursos de los sitios web?
   a) Ignorar completamente el archivo `robots.txt` de los sitios web.
   b) Visitar y rastrear todos los enlaces de una página web simultáneamente.
   c) Extraer únicamente contenido multimedia para reducir la carga en los servidores web.
   d) Seguir las directrices del archivo `robots.txt` y aplicar un retraso entre las solicitudes.

2. Considerando las dimensiones y desafíos inherentes a Big Data puede afirmarse, tomando en cuenta las características clave y las implicaciones para su procesamiento y análisis, que:
   a) Big Data se caracteriza principalmente por su pequeño volumen y uniformidad, permitiendo un procesamiento eficiente con mínimas adaptaciones de las herramientas de análisis de datos tradicionales.
   b) Big Data no desafía la capacidad de las herramientas tradicionales de procesamiento de datos para capturar, almacenar, gestionar y analizar efectivamente la información, dada la evolución constante de las capacidades computacionales y algoritmos de optimización.
   c) Aunque Big Data puede incluir datos estructurados, su naturaleza se expande al incorporar grandes cantidades de datos no estructurados y semiestructurados, lo que exige el uso de tecnologías especializadas en almacenamiento y procesamiento como Hadoop y sistemas de bases de datos NoSQL.
   d) Además de su complejidad y diversidad, Big Data introduce desafíos significativos en términos de veracidad y variabilidad al requerir métodos avanzados de limpieza y validación de datos para asegurar la integridad del análisis.

3. En un grafo una comunidad es:
   a) Un conjunto de nodos que no comparten ninguna similitud estructural o funcional entre sí.
   b) Un conjunto de nodos aislados.
   c) Un conjunto de nodos altamente conectados que forman un subgrafo completamente independiente del resto de la red.
   d) Un conjunto de nodos que están más densamente interconectados entre sí que con los nodos fuera del conjunto.

4. Se puede afirmar que:
   a) El término "Big Data" se refiere exclusivamente al volumen de datos que una organización maneja, sin tener en cuenta la velocidad, la variedad y la veracidad de los datos.
   b) Los SRI pueden beneficiarse de MapReduce para mejorar la RI relevante.
   c) MapReduce es un modelo de procesamiento distribuido utilizado para trabajar con grandes volúmenes de datos.
   d) Uno de los desafíos en el procesamiento de Big Data es la capacidad de gestionar y analizar datos provenientes de diversas fuentes y en diferentes formatos de manera eficiente.

5. En una biblioteca digital se necesita implementar un sistema que permita a los usuarios encontrar libros y artículos científicos de forma rápida y precisa. Los documentos están en diversos formatos, incluyendo PDF, EPUB y HTML. Se requiere seleccionar un algoritmo de indexación adecuado para el sistema, por lo que el programador designado para la implementación debe considerar:
   a) La velocidad de indexación y recuperación de datos.
   b) La capacidad para manejar documentos en diferentes formatos.
   c) La complejidad del algoritmo en términos de implementación y mantenimiento.
   d) La capacidad del algoritmo para procesar imágenes incrustadas o referenciadas en los ficheros.

6. Para contribuir positivamente al posicionamiento orgánico de un sitio web en los motores de búsqueda se puede:
   a) Crear contenido relevante y de alta calidad que satisfaga las necesidades de información de los usuarios.
   b) Mejorar la velocidad de carga del sitio web y asegurar que sea *responsive* y fácil de usar en dispositivos móviles.
   c) Incluir una densidad alta de palabras clave para asegurar que el sitio web aparezca en tantas búsquedas como sea posible.

d) Obtener enlaces entrantes de otros sitios web de alta autoridad y relevancia temática.

7. Si una red cumple la propiedad de ser un grafo de mundo pequeño, entonces se conoce que:
   a) El número de componentes fuertemente conexas está relacionado con la cantidad de grafos K_n presentes.
   b) La longitud media del camino entre todo par de vértices es pequeña.
   c) El grafo es un anillo regular de grado 5.
   d) La red posee pocos vértices.

8. ¿Qué algoritmos permiten obtener información de una red?
   a) Índices de centralidad.
   b) Detección de comunidades.
   c) Hypertext Induced Topic Selection (HITS).
   d) PageRank.

9. Sobre WordNet y su aplicación en el procesamiento del lenguaje natural, se puede afirmar que:
   a) El diseño de WordNet facilita su integración en aplicaciones multilingües de NLP, aunque su desarrollo original se centró en el inglés.
   b) Aunque WordNet es una herramienta valiosa en el NLP, su estructura no incluye información sobre la frecuencia de uso de las palabras en el lenguaje natural.
   c) Los synsets facilitan la identificación de relaciones semánticas entre palabras, como la hiperonimia y la meronimia, enriqueciendo tareas de NLP.
   d) WordNet proporciona una base para la desambiguación semántica al agrupar palabras con significados similares en synsets.

10. En el contexto del procesamiento de grandes conjuntos de datos, se puede asegurar que:
    a) MapReduce es un enfoque para procesar datos de forma secuencial en un solo servidor para evitar problemas de concurrencia.
    b) MapReduce divide una tarea en múltiples pasos de map y reduce que se ejecutan de forma secuencial en diferentes servidores para mejorar el rendimiento y la escalabilidad.
    c) MapReduce no es adecuado para el procesamiento de datos no estructurados.
    d) MapReduce solo puede manejar pequeñas cantidades de datos y no escala bien a grandes conjuntos de datos.

11. Si se tiene el conjunto de páginas interconectadas $\{A \to B, C; B \to C; C \to A\}$, entonces la página con valor más alto de PageRank es:
    a) A y B.
    b) A.
    c) C.
    d) B.

12. En el modelo de representación del conocimiento basado en herencia se puede asegurar que:
    a) La cantidad de padres que puede tener un nodo no es mayor que 4.
    b) Las conclusiones pueden ser canceladas si el grafo es ambiguo.
    c) Solo se usa en entornos referentes a la biología.
    d) El razonamiento deducido está respaldado por al menos un camino dentro del grafo.

13. Una plataforma de comercio electrónico desea mejorar la experiencia del usuario al permitir una navegación más personalizada y contextualizada. Actualmente los usuarios tienen dificultades para encontrar productos relevantes debido a la gran cantidad de opciones disponibles. La empresa está interesada en implementar características de la Web 2.5 y la Web Semántica para abordar este problema. ¿Qué características podrían ayudar para ofrecer una navegación más personalizada y contextualizada?
    a) La implementación de ontologías y metadatos para enriquecer la descripción de productos y mejorar la precisión de las recomendaciones.
    b) La optimización de la velocidad de carga del sitio web para mejorar la experiencia del usuario y reducir el abandono del carrito de compra.
    c) La integración de redes sociales para permitir la recomendación de productos basada en las preferencias de amigos y contactos.
    d) La incorporación de los productos en tendencia en el mercado.

14. ¿Qué es un "Uniform Resource Locator (URL) Frontier" en el contexto de Web Crawling?
    a) Un protocolo que define cómo se deben formatear las URLs para el crawling.
    b) Una lista prioritaria de URLs que aún no han sido visitadas por el crawler.
    c) Una técnica para filtrar URLs irrelevantes y mejorar la eficiencia del crawling.
    d) Una base de datos que almacena URLs únicas identificadas como recursos en la web.

15. En la RI en el contexto de Big Data, se puede asegurar que:
    a) La indexación distribuida divide los datos en múltiples fragmentos que se almacenan en varios nodos para permitir búsquedas paralelas y mejorar la escalabilidad.

b) La indexación distribuida no ofrece ventajas en términos de rendimiento y escalabilidad en comparación con la indexación centralizada.

c) La indexación distribuida es una técnica obsoleta.

d) La indexación distribuida almacena todos los datos en un solo servidor para facilitar su acceso y búsqueda.

16. La afirmación que mejor refleja el principio subyacente de PageRank, considerando su importancia en la clasificación de los sitios web, es:

a) PageRank valora más la cantidad de enlaces entrantes a una página web, independientemente de la calidad o relevancia de estos enlaces.

b) La efectividad de PageRank se basa exclusivamente en el análisis de las palabras clave contenidas en los enlaces entrantes, sin considerar la estructura de enlace de la Web.

c) El algoritmo de PageRank considera tanto la cantidad como la calidad de los enlaces entrantes, asignando mayor valor a los enlaces provenientes de sitios web considerados como "importantes".

d) PageRank opera bajo el supuesto de que los enlaces entrantes y salientes tienen el mismo impacto en la valoración de la relevancia de una página web.

17. La Web 2.0 se caracteriza por:

a) La existencia de sitios web dinámicos e interactivos que permiten a los usuarios participar, comentar e interactuar tanto con los creadores de contenido como con otros usuarios.

b) El uso de distintas tecnologías para crear experiencias web más interactivas y con mayor capacidad de respuesta.

c) Las actualizaciones del contenido de los sitios web se efectúa de forma manual dentro del código HTML

d) Las plataformas ofrecen experiencias a la medida, permitiendo a los usuarios personalizar sus perfiles, recibir recomendaciones ajustadas al contenido y participar en filtrado colaborativo.

18. La centralidad de intermediación de un nodo indica:

a) La cantidad de nodos vecinos directos.

b) La resistencia del nodo a fallos.

c) El número total de conexiones entrantes y salientes.

d) La frecuencia con la que un nodo actúa como puente en el camino más corto entre otros dos nodos.

19. Los algoritmos para detectar comunidades en una red intentan:

a) Buscar conjuntos donde cada nodo de un mismo conjunto tenga características similares al resto de los nodos del conjunto.

b) Seleccionar aleatoriamente nodos de alto grado y sus vecinos.

c) Buscar subgrafos tal que no incluyan nodos cuya ausencia desconecte al subgrafo.

d) Encontrar grupos donde los nodos pertenecientes a los mismos grupos son cercanos bajo cierta métrica y lejanos con respecto a los nodos de otros grupos.

20. Considerando las prácticas éticas y legales en el Web Scraping, se puede asegurar que:

a) Web Scraping requiere considerar las políticas de `robots.txt` del sitio web objetivo.

b) Es importante revisar y respetar los términos de servicio del sitio web, así como las leyes aplicables de protección de datos y derechos de autor, antes de realizar Web Scraping.

c) Web Scraping sobre datos personales sin consentimiento es generalmente aceptado si los datos se utilizan con fines de investigación.

d) La extracción de datos mediante Web Scraping siempre es legal, independientemente de las leyes locales sobre derechos de autor y privacidad de datos definidos en los sitios web.

Trabajo de Control
Sistemas de Recuperación de Información
Fecha: 15 de julio de 2024
Temario #7

Nombre: _____ Grupo: _____

Seleccione en cada caso la(s) respuesta(s) correcta(s). Considere también no marcar ninguna opción.

1. Para la RI, el análisis de las redes puede:
   a) Ayudar a identificar grupos de interés.
   b) Indicar la importancia de una entidad en la transmisión de la información.
   c) Ayudar a comprender la conectividad y la accesibilidad entre las entidades.
   d) Revelar patrones de influencia dentro de una comunidad.

2. El conjunto frontera de URLs en un Web Crawler:
   a) Delimita los sitios web que visitará en futuras iteraciones del proceso.
   b) Es similar a un conjunto de URLs que esperan ser visitadas.
   c) Almacena hipervínculos que pertenecen al mismo dominio del conjunto semilla de URLs con que inició el crawler.
   d) Indica las secciones que no pueden ser visitadas de cada sitio web.

3. ¿Cuál de las siguientes opciones describe mejor la diferencia clave entre Web Crawling y Web Scraping?
   a) Web Crawling se centra en la exploración y recopilación de enlaces de múltiples sitios web, mientras que Web Scraping se enfoca en la extracción específica de datos de páginas web individuales.
   b) Web Crawling se realiza utilizando herramientas de automatización como Selenium WebDriver, mientras que Web Scraping se lleva a cabo mediante el análisis de HTML y CSS.
   c) Web Scraping implica el análisis de la estructura y el contenido de las páginas web para extraer datos, mientras que Web Crawling se refiere a la descarga y almacenamiento de páginas web completas.
   d) Web Scraping es más eficaz para rastrear e indexar contenido web para motores de búsqueda, mientras que Web Crawling se utiliza principalmente para la extracción de datos en proyectos de investigación.

4. Referente al Web Crawler puede afirmarse que:
   a) Los hipervínculos encontrados en cada sitio web que no pertenecen al dominio donde fueron encontrados se desechan, puesto que no expande el conjunto de URLs sin visitar.
   b) Las páginas visitadas no se procesan nunca más.
   c) No necesita de un conjunto inicial de URLs para recorrer la Web.
   d) No tiene como objetivo indexar y recopilar información de diferentes sitios web.

5. En el contexto de la representación del conocimiento basada en herencia, ¿qué caracteriza a la herencia cancelable?
   a) La herencia es el resultado del razonamiento no transitivo.
   b) Las conclusiones no están determinadas y dependen del nodo de interés.
   c) Las propiedades heredadas siempre se mantienen y no pueden anularse.
   d) No existe ambigüedad en las conclusiones obtenidas.

6. La transición de la Web 1.0 a la Web 2.0 se caracterizó principalmente por:
   a) El aumento en la velocidad de conexión a internet, que permitió una mejor calidad de las páginas web.
   b) La disminución de la importancia de los motores de búsqueda en la navegación web.
   c) El cambio de páginas web estáticas a dinámicas, permitiendo la interacción del usuario y la generación de contenido.
   d) La reducción en el uso de HTML y CSS en el desarrollo de sitios web.

7. La política de ordenación de URLs en los Web Crawlers tiene como aspecto fundamental:
   a) Limitar el acceso de los crawlers a ciertas secciones de un sitio web, evitando el rastreo de URLs consideradas menos importantes o sensibles.
   b) Definir la estructura de la URL de destino, asegurando que estén ordenadas alfabéticamente para facilitar la navegación y la indexación.
   c) Establecer la prioridad de rastreo de las URLs, determinando el orden en que los crawlers visitan y procesan cada página web.
   d) Determinar la forma en que los crawlers asignan un valor de relevancia a cada URL para clasificarlas en el índice de búsqueda.

8. La computación evolutiva:

a) No es aplicable en la RI debido a la complejidad de los algoritmos evolutivos.

b) Solo puede manejar conjuntos de datos pequeños y no es escalable a grandes volúmenes de datos.

c) Utiliza algoritmos para buscar soluciones óptimas en grandes espacios de búsqueda, lo que la hace adecuada para problemas de optimización en la RI.

d) Es útil solo para problemas de clasificación de documentos y no para otras tareas de RI en general.

9. La afirmación que mejor refleja el principio subyacente de PageRank, considerando su importancia en la clasificación de los sitios web, es:

a) PageRank valora más la cantidad de enlaces entrantes a una página web, independientemente de la calidad o relevancia de estos enlaces.

b) La efectividad de PageRank se basa exclusivamente en el análisis de las palabras clave contenidas en los enlaces entrantes, sin considerar la estructura de enlace de la Web.

c) El algoritmo de PageRank considera tanto la cantidad como la calidad de los enlaces entrantes, asignando mayor valor a los enlaces provenientes de sitios web considerados como "importantes".

d) PageRank opera bajo el supuesto de que los enlaces entrantes y salientes tienen el mismo impacto en la valoración de la relevancia de una página web.

10. La integración de Hadoop y MapReduce en la RI trae como ventaja:

a) La eliminación de la necesidad de sistemas de bases de datos.

b) La posibilidad de la extracción de información relevante y la generación de resultados significativos de grandes conjuntos de datos.

c) La garantía de la privacidad absoluta de los datos procesados.

d) La reducción de los costos operativos a cero.

11. Se tiene un grafo donde cada nodo es un personaje de cierto libro de cuentos y la existencia de las aristas está definida si dos personajes aparecen en el mismo cuento. Se puede asegurar que:

a) La centralidad de intermediación identifica a los personajes que actúan como conectores entre personajes de cuentos distintos.

b) La centralidad de grado es útil para identificar los personajes con más conexiones dentro de la red, lo que puede indicar su importancia en el libro.

c) La centralidad de cercanía ofrece una relación entre la cantidad de vecinos de un nodo con respecto a la longitud máxima de un camino dentro del grafo partiendo del nodo en cuestión.

d) La centralidad de cercanía indica el grado de conexión de cada personaje con el resto de los personajes de los cuentos del libro.

12. En un grafo una comunidad es:

a) Un conjunto de nodos que no comparten ninguna similitud estructural o funcional entre sí.

b) Un conjunto de nodos aislados.

c) Un conjunto de nodos altamente conectados que forman un subgrafo completamente independiente del resto de la red.

d) Un conjunto de nodos que están más densamente interconectados entre sí que con los nodos fuera del conjunto.

13. La premisa básica del algoritmo de PageRank para clasificar páginas web en los resultados de búsqueda es:

a) La antigüedad de la página web es el principal factor para determinar su clasificación.

b) El contenido y la relevancia de las palabras clave en la página web determinan su posición.

c) Los enlaces entrantes a una página web desde otras páginas contribuyen a su importancia y clasificación.

d) La cantidad de visitas que recibe una página web determina su clasificación.

14. Una empresa de comercio electrónico necesita mejorar su motor de búsqueda para proporcionar resultados más relevantes a sus usuarios. Actualmente, los resultados de la búsqueda no son precisos y los usuarios a menudo encuentran dificultades para encontrar productos específicos. La empresa está considerando implementar una indexación por conceptos para mejorar la relevancia de los resultados de búsqueda. ¿Qué beneficios podría provocar este cambio?

a) Permite adaptarse fácilmente a cambios en el vocabulario y la terminología utilizada en las descripciones de los productos.

b) Ayuda a identificar automáticamente relaciones entre productos, lo que puede mejorar las recomendaciones personalizadas a los usuarios.

c) Facilita la visualización de productos al agruparlos por categorías o características comunes.

d) Mejora la precisión en la búsqueda de productos relacionados, incluso cuando no coinciden exactamente con los términos de búsqueda del usuario.

15. En una biblioteca digital se necesita implementar un sistema que permita a los usuarios encontrar libros y artículos científicos de forma rápida y precisa. Los documentos están en diversos formatos, incluyendo PDF, EPUB y HTML. Se requiere seleccionar un algoritmo de indexación adecuado para el sistema, por lo que el programador designado para la implementación debe considerar:

a) La velocidad de indexación y recuperación de datos.

b) La capacidad para manejar documentos en diferentes formatos.

c) La complejidad del algoritmo en términos de implementación y mantenimiento.

d) La capacidad del algoritmo para procesar imágenes incrustadas o referenciadas en los ficheros.

16. Para contribuir positivamente al posicionamiento orgánico de un sitio web en los motores de búsqueda se puede:

a) Crear contenido relevante y de alta calidad que satisfaga las necesidades de información de los usuarios.

b) Mejorar la velocidad de carga del sitio web y asegurar que sea *responsive* y fácil de usar en dispositivos móviles.

c) Incluir una densidad alta de palabras clave para asegurar que el sitio web aparezca en tantas búsquedas como sea posible.

d) Obtener enlaces entrantes de otros sitios web de alta autoridad y relevancia temática.

17. En el algoritmo de Indexación Basada en Clasificación Bloqueada (BSBI), ¿cuál es el paso final para crear un índice invertido para la colección completa de los datos?

a) Eliminar los términos duplicados de los índices.

b) Indexar cada bloque de forma independiente.

c) Fusionar los índices invertidos de cada bloque.

d) Dividir la colección de datos en bloques de tamaño fijo.

18. No se considera como técnica para detectar comunidades en una red:

a) Analizar la mutualidad de los enlaces.

b) Usar el agrupamiento jerárquico.

c) Utilizar el algoritmo de K-Means.

d) Encontrar cliques de vértices de grado par.

19. Considerando las dimensiones y desafíos inherentes a Big Data puede afirmarse, tomando en cuenta las características clave y las implicaciones para su procesamiento y análisis, que:

a) Big Data se caracteriza principalmente por su pequeño volumen y uniformidad, permitiendo un procesamiento eficiente con mínimas adaptaciones de las herramientas de análisis de datos tradicionales.

b) Big Data no desafía la capacidad de las herramientas tradicionales de procesamiento de datos para capturar, almacenar, gestionar y analizar efectivamente la información, dada la evolución constante de las capacidades computacionales y algoritmos de optimización.

c) Aunque Big Data puede incluir datos estructurados, su naturaleza se expande al incorporar grandes cantidades de datos no estructurados y semiestructurados, lo que exige el uso de tecnologías especializadas en almacenamiento y procesamiento como Hadoop y sistemas de bases de datos NoSQL.

d) Además de su complejidad y diversidad, Big Data introduce desafíos significativos en términos de veracidad y variabilidad al requerir métodos avanzados de limpieza y validación de datos para asegurar la integridad del análisis.

20. Sobre el algoritmo de PageRank, visto en clase, se puede afirmar que:

a) Evalúa la importancia de un sitio web en función de la calidad y cantidad de enlaces entrantes que recibe de otros sitios web.

b) Solo tiene en cuenta el contenido en un sitio web para determinar su relevancia en los resultados de búsqueda.

c) Asigna una puntuación alta a los sitios web que tienen un gran número de enlaces entrantes sin tener en cuenta la calidad de esos enlaces.

d) Asigna una puntuación baja a los sitios web que contienen muchos enlaces salientes, ya que indica una falta de relevancia.

Trabajo de Control
Sistemas de Recuperación de Información
Fecha: 15 de julio de 2024
Temario #8

Nombre: _____ Grupo: _____

Seleccione en cada caso la(s) respuesta(s) correcta(s). Considere también no marcar ninguna opción.

1. En un SRI la indexación:
   a) Consiste en asociar un identificador único a cada dato almacenado en el sistema.
   b) Mejora la experiencia del usuario.
   c) Permite la organización y la categorización de la información.
   d) Facilita la RI relevante.

2. El algoritmo de PageRank converge si:
   a) La norma de la diferencia entre los vectores es menor a un umbral predefinido.
   b) Finaliza la ejecución cuando el número de iteraciones excede un máximo de iteraciones previamente definido.
   c) Se define un factor de normalización en la fórmula de la función.
   d) El algoritmo no se implementa de forma iterativa.

3. En una red de transporte donde cada nodo es una parada de autobús y las aristas representan si existe un carro que pasa por ambos sitios, ¿qué puede mejorar el sistema de transporte?
   a) No considerar la centralidad de intermediación de las estaciones de transporte público al planificar rutas y horarios, ya que no tiene impacto en las conexiones entre las paradas.
   b) Utilizar el grafo inducido de los nodos con mayor valor en la centralidad de grado para aplicar la centralidad de intermediación con el propósito de reforzar las paradas con mayor tráfico.
   c) Utilizar la centralidad de intermediación para identificar las paradas de transferencia clave y establecer nuevas conexiones entre diferentes líneas de transporte público.
   d) No implementar sistemas de información en tiempo real para los usuarios, ya que pueden aumentar la carga de los trabajadores.

4. Una empresa de comercio electrónico necesita mejorar su motor de búsqueda para proporcionar resultados más relevantes a sus usuarios. Actualmente, los resultados de la búsqueda no son precisos y los usuarios a menudo encuentran dificultades para encontrar productos específicos. La empresa está considerando implementar una indexación por conceptos para mejorar la relevancia de los resultados de búsqueda. ¿Qué beneficios podría provocar este cambio?
   a) Permite adaptarse fácilmente a cambios en el vocabulario y la terminología utilizada en las descripciones de los productos.
   b) Ayuda a identificar automáticamente relaciones entre productos, lo que puede mejorar las recomendaciones personalizadas a los usuarios.
   c) Facilita la visualización de productos al agruparlos por categorías o características comunes.
   d) Mejora la precisión en la búsqueda de productos relacionados, incluso cuando no coinciden exactamente con los términos de búsqueda del usuario.

5. Los algoritmos para detectar comunidades en una red intentan:
   a) Buscar conjuntos donde cada nodo de un mismo conjunto tenga características similares al resto de los nodos del conjunto.
   b) Seleccionar aleatoriamente nodos de alto grado y sus vecinos.
   c) Buscar subgrafos tal que no incluyan nodos cuya ausencia desconecte al subgrafo.
   d) Encontrar grupos donde los nodos pertenecientes a los mismos grupos son cercanos bajo cierta métrica y lejanos con respecto a los nodos de otros grupos.

6. Para contribuir positivamente al posicionamiento orgánico de un sitio web en los motores de búsqueda se puede:
   a) Crear contenido relevante y de alta calidad que satisfaga las necesidades de información de los usuarios.
   b) Mejorar la velocidad de carga del sitio web y asegurar que sea *responsive* y fácil de usar en dispositivos móviles.
   c) Incluir una densidad alta de palabras clave para asegurar que el sitio web aparezca en tantas búsquedas como sea posible.
   d) Obtener enlaces entrantes de otros sitios web de alta autoridad y relevancia temática.

7. El componente responsable de la gestión de recursos y planificación de tareas en Hadoop es:

a) MapReduce.

b) Hadoop Common.

c) HDFS.

d) YARN.

8. En el contexto de la RI en redes, se puede afirmar que:

   a) La detección de comunidades no es relevante para la RI en redes, ya que se centra únicamente en la estructura de la red sin considerar el contenido.

   b) La detección de comunidades en una red siempre produce resultados objetivos y consistentes independientemente del algoritmo utilizado.

   c) La detección de comunidades ayuda a identificar grupos de nodos altamente conectados entre sí, lo que puede ser útil para comprender la estructura y el contenido de la red.

   d) La detección de comunidades solo se aplica a redes pequeñas y simples, no a redes grandes y complejas.

9. La afirmación que mejor describe la política de amabilidad en los Web Crawlers es:

   a) Los crawlers se diseñan para acceder a sitios web sin restricciones y extraer datos de manera agresiva para su indexación.

   b) La política de amabilidad de los Web Crawlers dicta que los crawlers deben priorizar ciertos tipos de contenido sobre otros, ignorando completamente ciertas páginas web.

   c) La política de amabilidad establece pautas y reglas sobre cómo los crawlers deben interactuar con los sitios web para minimizar la carga del servidor y respetar las directivas de los administradores del sitio.

   d) Los Web Crawlers son libres de recopilar datos de cualquier sitio web sin restricciones, independientemente de la cantidad de tráfico que generen.

10. Una plataforma de comercio electrónico desea mejorar la experiencia del usuario al permitir una navegación más personalizada y contextualizada. Actualmente los usuarios tienen dificultades para encontrar productos relevantes debido a la gran cantidad de opciones disponibles. La empresa está interesada en implementar características de la Web 2.5 y la Web Semántica para abordar este problema. ¿Qué características podrían ayudar para ofrecer una navegación más personalizada y contextualizada?

    a) La implementación de ontologías y metadatos para enriquecer la descripción de productos y mejorar la precisión de las recomendaciones.

    b) La optimización de la velocidad de carga del sitio web para mejorar la experiencia del usuario y reducir el abandono del carrito de compra.

    c) La integración de redes sociales para permitir la recomendación de productos basada en las preferencias de amigos y contactos.

    d) La incorporación de los productos en tendencia en el mercado.

11. Se tiene un grafo donde cada nodo es un personaje de cierto libro de cuentos y la existencia de las aristas está definida si dos personajes aparecen en el mismo cuento. Se puede asegurar que:

    a) La centralidad de intermediación identifica a los personajes que actúan como conectores entre personajes de cuentos distintos.

    b) La centralidad de grado es útil para identificar los personajes con más conexiones dentro de la red, lo que puede indicar su importancia en el libro.

    c) La centralidad de cercanía ofrece una relación entre la cantidad de vecinos de un nodo con respecto a la longitud máxima de un camino dentro del grafo partiendo del nodo en cuestión.

    d) La centralidad de cercanía indica el grado de conexión de cada personaje con el resto de los personajes de los cuentos del libro.

12. Si una red cumple la propiedad de ser un grafo de mundo pequeño, entonces se conoce que:

    a) El número de componentes fuertemente conexas está relacionado con la cantidad de grafos K_n presentes.

    b) La longitud media del camino entre todo par de vértices es pequeña.

    c) El grafo es un anillo regular de grado 5.

    d) La red posee pocos vértices.

13. Si se tiene el conjunto de páginas interconectadas $\{A \to B, C; B \to C; C \to A\}$, entonces la página con valor más alto de PageRank es:

    a) A y B.

    b) A.

    c) C.

    d) B.

14. En el diseño de un videojuego de roles (RPG) en el que los personajes tienen atributos como salud, fuerza y velocidad, ¿cuál de las siguientes opciones representa mejor una implementación de la representación del conocimiento orientado a objetos?

    a) Cada personaje se representa como una lista de cadenas de texto que describen sus características físicas y habilidades.

    b) Cada personaje se representa como una función que calcula sus atributos en función de su nivel y experiencia.

    c) Cada personaje se representa como una matriz de números que almacena sus valores de atributos.

d) Cada personaje se representa como un objeto con propiedades como salud, fuerza y velocidad, y métodos para modificar y consultar estos valores.

15. En la RI en el contexto de Big Data, se puede asegurar que:
    a) El almacenamiento distribuido centraliza todos los datos en un único servidor para facilitar su gestión y mantenimiento.
    b) El almacenamiento distribuido ofrece ventajas significativas en términos de escalabilidad y rendimiento en comparación con el almacenamiento centralizado.
    c) El almacenamiento distribuido reparte los datos en múltiples servidores para mejorar la disponibilidad y la redundancia del sistema.
    d) El almacenamiento distribuido es una técnica obsoleta.

16. La centralidad de intermediación de un nodo indica:
    a) La cantidad de nodos vecinos directos.
    b) La resistencia del nodo a fallos.
    c) El número total de conexiones entrantes y salientes.
    d) La frecuencia con la que un nodo actúa como puente en el camino más corto entre otros dos nodos.

17. La Web 3 se conoce como:
    a) Internet de las cosas.
    b) Web de solo lectura.
    c) Web semántica.
    d) Web de escritura-lectura.

18. El algoritmo de PageRank puede describirse como un procedimiento utilizado para:
    a) Determinar la velocidad de carga de una página web en un navegador.
    b) Calcular la relevancia de una página web en función de la cantidad y calidad de los enlaces que apuntan hacia ella.
    c) Prevenir el spam y el contenido no deseado en las páginas web.
    d) Clasificar las páginas web en función de su edad y autoridad.

19. Una buena práctica de SEO para mejorar el posicionamiento de un sitio web en los motores de búsqueda es:
    a) Obtener enlaces de otros sitios web relevantes y de calidad que apunten al sitio.
    b) Copiar contenido directamente de otros sitios web populares para aumentar la cantidad de páginas indexadas.
    c) Utilizar técnicas de encubrimiento para mostrar contenido diferente a lo indexado por los motores de búsqueda.
    d) Llenar el contenido con palabras clave irrelevantes para aumentar la densidad de palabras clave.

20. Considerando las dimensiones y desafíos inherentes a Big Data puede afirmarse, tomando en cuenta las características clave y las implicaciones para su procesamiento y análisis, que:
    a) Big Data se caracteriza principalmente por su pequeño volumen y uniformidad, permitiendo un procesamiento eficiente con mínimas adaptaciones de las herramientas de análisis de datos tradicionales.
    b) Big Data no desafía la capacidad de las herramientas tradicionales de procesamiento de datos para capturar, almacenar, gestionar y analizar efectivamente la información, dada la evolución constante de las capacidades computacionales y algoritmos de optimización.
    c) Aunque Big Data puede incluir datos estructurados, su naturaleza se expande al incorporar grandes cantidades de datos no estructurados y semiestructurados, lo que exige el uso de tecnologías especializadas en almacenamiento y procesamiento como Hadoop y sistemas de bases de datos NoSQL.
    d) Además de su complejidad y diversidad, Big Data introduce desafíos significativos en términos de veracidad y variabilidad al requerir métodos avanzados de limpieza y validación de datos para asegurar la integridad del análisis.

Trabajo de Control
Sistemas de Recuperación de Información
Fecha: 15 de julio de 2024
Temario #9

Nombre: _____ Grupo: _____

Seleccione en cada caso la(s) respuesta(s) correcta(s). Considere también no marcar ninguna opción.

1. Se puede afirmar que:
   a) No es posible identificar subgrupos dentro de una red utilizando análisis de redes.
   b) La cantidad de conexiones de un nodo siempre indica su influencia en la red.
   c) El tamaño de una red es siempre indicativo de su efectividad en la transmisión de información.
   d) Todas las relaciones en una red tienen la misma importancia para el análisis.

2. En un sistema donde el conocimiento está orientado a objetos se puede asegurar que:
   a) Se enfatiza la atención a la información de la cual se extrajo el conocimiento.
   b) Los marcos y las bandas son las estructuras utilizadas para la representación del modelo.
   c) No es posible definir especializaciones de los objetos de la vida real dentro del sistema.
   d) Las funciones de agregación dificultan poder establecer relaciones entre los objetos.

3. Considerando las dimensiones y desafíos inherentes a Big Data puede afirmarse, tomando en cuenta las características clave y las implicaciones para su procesamiento y análisis, que:
   a) Big Data se caracteriza principalmente por su pequeño volumen y uniformidad, permitiendo un procesamiento eficiente con mínimas adaptaciones de las herramientas de análisis de datos tradicionales.
   b) Big Data no desafía la capacidad de las herramientas tradicionales de procesamiento de datos para capturar, almacenar, gestionar y analizar efectivamente la información, dada la evolución constante de las capacidades computacionales y algoritmos de optimización.
   c) Aunque Big Data puede incluir datos estructurados, su naturaleza se expande al incorporar grandes cantidades de datos no estructurados y semiestructurados, lo que exige el uso de tecnologías especializadas en almacenamiento y procesamiento como Hadoop y sistemas de bases de datos NoSQL.
   d) Además de su complejidad y diversidad, Big Data introduce desafíos significativos en términos de veracidad y variabilidad al requerir métodos avanzados de limpieza y validación de datos para asegurar la integridad del análisis.

4. Sobre WordNet y su aplicación en el procesamiento del lenguaje natural, se puede afirmar que:
   a) El diseño de WordNet facilita su integración en aplicaciones multilingües de NLP, aunque su desarrollo original se centró en el inglés.
   b) Aunque WordNet es una herramienta valiosa en el NLP, su estructura no incluye información sobre la frecuencia de uso de las palabras en el lenguaje natural.
   c) Los synsets facilitan la identificación de relaciones semánticas entre palabras, como la hiperonimia y la meronimia, enriqueciendo tareas de NLP.
   d) WordNet proporciona una base para la desambiguación semántica al agrupar palabras con significados similares en synsets.

5. Una buena práctica de SEO para mejorar el posicionamiento de un sitio web en los motores de búsqueda es:
   a) Obtener enlaces de otros sitios web relevantes y de calidad que apunten al sitio.
   b) Copiar contenido directamente de otros sitios web populares para aumentar la cantidad de páginas indexadas.
   c) Utilizar técnicas de encubrimiento para mostrar contenido diferente a lo indexado por los motores de búsqueda.
   d) Llenar el contenido con palabras clave irrelevantes para aumentar la densidad de palabras clave.

6. Referente al Web Crawler puede afirmarse que:
   a) Los hipervínculos encontrados en cada sitio web que no pertenecen al dominio donde fueron encontrados se desechan, puesto que no expande el conjunto de URLs sin visitar.
   b) Las páginas visitadas no se procesan nunca más.
   c) No necesita de un conjunto inicial de URLs para recorrer la Web.

d) No tiene como objetivo indexar y recopilar información de diferentes sitios web.

7. Implementar índices invertidos en un SRI asegura:
   a) Acelerar el proceso de búsqueda al permitir búsquedas directas por contenido en lugar de por título.
   b) Reducir la cantidad de espacio de almacenamiento necesario al comprimir los datos de los documentos.
   c) Facilitar la búsqueda sobre los datos que contienen términos específicos al mantener una lista de datos para cada término único.
   d) Incrementar la seguridad de los datos almacenados al dificultar el acceso directo a la información sin el índice correcto.

8. El posicionamiento de un sitio web en los motores de búsqueda puede ser afectado por:
   a) Utilizar técnicas de encubrimiento para mostrar contenido diferente a lo indexado por los motores de búsqueda.
   b) Contenido sin valor en el sitio web.
   c) Mantener una estructura de URL clara y coherente.
   d) Obtener enlaces de sitios web irrelevantes y de baja calidad.

9. La afirmación que mejor describe la política de amabilidad en los Web Crawlers es:
   a) Los crawlers se diseñan para acceder a sitios web sin restricciones y extraer datos de manera agresiva para su indexación.
   b) La política de amabilidad de los Web Crawlers dicta que los crawlers deben priorizar ciertos tipos de contenido sobre otros, ignorando completamente ciertas páginas web.
   c) La política de amabilidad establece pautas y reglas sobre cómo los crawlers deben interactuar con los sitios web para minimizar la carga del servidor y respetar las directivas de los administradores del sitio.
   d) Los Web Crawlers son libres de recopilar datos de cualquier sitio web sin restricciones, independientemente de la cantidad de tráfico que generen.

10. En un grafo una comunidad es:
    a) Un conjunto de nodos que no comparten ninguna similitud estructural o funcional entre sí.
    b) Un conjunto de nodos aislados.
    c) Un conjunto de nodos altamente conectados que forman un subgrafo completamente independiente del resto de la red.
    d) Un conjunto de nodos que están más densamente interconectados entre sí que con los nodos fuera del conjunto.

11. Considerando las prácticas éticas y legales en el Web Scraping, se puede asegurar que:
    a) Web Scraping requiere considerar las políticas de `robots.txt` del sitio web objetivo.
    b) Es importante revisar y respetar los términos de servicio del sitio web, así como las leyes aplicables de protección de datos y derechos de autor, antes de realizar Web Scraping.
    c) Web Scraping sobre datos personales sin consentimiento es generalmente aceptado si los datos se utilizan con fines de investigación.
    d) La extracción de datos mediante Web Scraping siempre es legal, independientemente de las leyes locales sobre derechos de autor y privacidad de datos definidos en los sitios web.

12. En el contexto del procesamiento de grandes conjuntos de datos, se puede asegurar que:
    a) MapReduce es un enfoque para procesar datos de forma secuencial en un solo servidor para evitar problemas de concurrencia.
    b) MapReduce divide una tarea en múltiples pasos de map y reduce que se ejecutan de forma secuencial en diferentes servidores para mejorar el rendimiento y la escalabilidad.
    c) MapReduce no es adecuado para el procesamiento de datos no estructurados.
    d) MapReduce solo puede manejar pequeñas cantidades de datos y no escala bien a grandes conjuntos de datos.

13. El algoritmo de PageRank converge si:
    a) La norma de la diferencia entre los vectores es menor a un umbral predefinido.
    b) Finaliza la ejecución cuando el número de iteraciones excede un máximo de iteraciones previamente definido.
    c) Se define un factor de normalización en la fórmula de la función.
    d) El algoritmo no se implementa de forma iterativa.

14. Si una red cumple la propiedad de ser un grafo de mundo pequeño, entonces se conoce que:
    a) El número de componentes fuertemente conexas está relacionado con la cantidad de grafos K_n presentes.
    b) La longitud media del camino entre todo par de vértices es pequeña.
    c) El grafo es un anillo regular de grado 5.
    d) La red posee pocos vértices.

15. La política de ordenación de URLs en los Web Crawlers tiene como aspecto fundamental:

a) Limitar el acceso de los crawlers a ciertas secciones de un sitio web, evitando el rastreo de URLs consideradas menos importantes o sensibles.

b) Definir la estructura de la URL de destino, asegurando que estén ordenadas alfabéticamente para facilitar la navegación y la indexación.

c) Establecer la prioridad de rastreo de las URLs, determinando el orden en que los crawlers visitan y procesan cada página web.

d) Determinar la forma en que los crawlers asignan un valor de relevancia a cada URL para clasificarlas en el índice de búsqueda.

16. Se puede afirmar que:

a) El término "Big Data" se refiere exclusivamente al volumen de datos que una organización maneja, sin tener en cuenta la velocidad, la variedad y la veracidad de los datos.

b) Los SRI pueden beneficiarse de MapReduce para mejorar la RI relevante.

c) MapReduce es un modelo de procesamiento distribuido utilizado para trabajar con grandes volúmenes de datos.

d) Uno de los desafíos en el procesamiento de Big Data es la capacidad de gestionar y analizar datos provenientes de diversas fuentes y en diferentes formatos de manera eficiente.

17. En el contexto de la RI en redes sociales se puede afirmar que:

a) El análisis de redes sociales se centra exclusivamente en la cantidad de seguidores que tiene un usuario en particular para determinar su influencia en la red.

b) El análisis de centralidad de intermediación se utiliza para identificar usuarios que son importantes en una red social debido a su posición como "puentes" entre diferentes grupos de usuarios.

c) El análisis de sentimientos se utiliza para determinar la popularidad de una publicación en redes sociales sin tener en cuenta la opinión de los usuarios.

d) El análisis de redes sociales no es útil para comprender la difusión de información en una red social específica, para ello se utiliza la medida de centralidad de vector propio.

18. El algoritmo Hypertext Induced Topic Selection (HITS) intenta buscar nodos especiales. Estos son conocidos como:

a) Autoridades.

b) Hubs.

c) Centrales.

d) Sensibles.

19. La computación evolutiva:

a) No es aplicable en la RI debido a la complejidad de los algoritmos evolutivos.

b) Solo puede manejar conjuntos de datos pequeños y no es escalable a grandes volúmenes de datos.

c) Utiliza algoritmos para buscar soluciones óptimas en grandes espacios de búsqueda, lo que la hace adecuada para problemas de optimización en la RI.

d) Es útil solo para problemas de clasificación de documentos y no para otras tareas de RI en general.

20. Los algoritmos para detectar comunidades en una red intentan:

a) Buscar conjuntos donde cada nodo de un mismo conjunto tenga características similares al resto de los nodos del conjunto.

b) Seleccionar aleatoriamente nodos de alto grado y sus vecinos.

c) Buscar subgrafos tal que no incluyan nodos cuya ausencia desconecte al subgrafo.

d) Encontrar grupos donde los nodos pertenecientes a los mismos grupos son cercanos bajo cierta métrica y lejanos con respecto a los nodos de otros grupos.

Trabajo de Control
Sistemas de Recuperación de Información
Fecha: 15 de julio de 2024
Temario #10

Nombre: _____ Grupo: _____

Seleccione en cada caso la(s) respuesta(s) correcta(s). Considere también no marcar ninguna opción.

1. Se tiene un grafo donde cada nodo es un personaje de cierto libro de cuentos y la existencia de las aristas está definida si dos personajes aparecen en el mismo cuento. Se puede asegurar que:
   a) La centralidad de intermediación identifica a los personajes que actúan como conectores entre personajes de cuentos distintos.
   b) La centralidad de grado es útil para identificar los personajes con más conexiones dentro de la red, lo que puede indicar su importancia en el libro.
   c) La centralidad de cercanía ofrece una relación entre la cantidad de vecinos de un nodo con respecto a la longitud máxima de un camino dentro del grafo partiendo del nodo en cuestión.
   d) La centralidad de cercanía indica el grado de conexión de cada personaje con el resto de los personajes de los cuentos del libro.

2. La afirmación que mejor refleja el principio subyacente de PageRank, considerando su importancia en la clasificación de los sitios web, es:
   a) PageRank valora más la cantidad de enlaces entrantes a una página web, independientemente de la calidad o relevancia de estos enlaces.
   b) La efectividad de PageRank se basa exclusivamente en el análisis de las palabras clave contenidas en los enlaces entrantes, sin considerar la estructura de enlace de la Web.
   c) El algoritmo de PageRank considera tanto la cantidad como la calidad de los enlaces entrantes, asignando mayor valor a los enlaces provenientes de sitios web considerados como "importantes".
   d) PageRank opera bajo el supuesto de que los enlaces entrantes y salientes tienen el mismo impacto en la valoración de la relevancia de una página web.

3. La integración de Hadoop y MapReduce en la RI trae como ventaja:
   a) La eliminación de la necesidad de sistemas de bases de datos.
   b) La posibilidad de la extracción de información relevante y la generación de resultados significativos de grandes conjuntos de datos.
   c) La garantía de la privacidad absoluta de los datos procesados.
   d) La reducción de los costos operativos a cero.

4. El algoritmo de PageRank converge si:
   a) La norma de la diferencia entre los vectores es menor a un umbral predefinido.
   b) Finaliza la ejecución cuando el número de iteraciones excede un máximo de iteraciones previamente definido.
   c) Se define un factor de normalización en la fórmula de la función.
   d) El algoritmo no se implementa de forma iterativa.

5. En un sistema de control de tráfico urbano basado en reglas, ¿cuál de las siguientes reglas sería más efectiva para manejar situaciones de congestión en una intersección?
   a) Si hay pocos vehículos en la intersección, reducir el tiempo de los semáforos en verde.
   b) Si hay muchos vehículos en la intersección, aumentar el tiempo de los semáforos en verde.
   c) Si hay un vehículo de emergencia en la intersección, detener todos los demás vehículos.
   d) Si hay muchos peatones cruzando la intersección, reducir el tiempo de los semáforos en rojo.

6. Analizar una red permite:
   a) Detectar posibles tendencias antes de que se conviertan en tendencia.
   b) Obtener predicciones exactas de eventos futuros en mercados financieros.
   c) Evaluar la calidad del contenido de un sitio web.
   d) Encontrar nodos "sensibles" o críticos para la red.

7. La afirmación que mejor describe la política de amabilidad en los Web Crawlers es:
   a) Los crawlers se diseñan para acceder a sitios web sin restricciones y extraer datos de manera agresiva para su indexación.
   b) La política de amabilidad de los Web Crawlers dicta que los crawlers deben priorizar ciertos tipos de contenido sobre otros, ignorando completamente ciertas páginas web.

c) La política de amabilidad establece pautas y reglas sobre cómo los crawlers deben interactuar con los sitios web para minimizar la carga del servidor y respetar las directivas de los administradores del sitio.

d) Los Web Crawlers son libres de recopilar datos de cualquier sitio web sin restricciones, independientemente de la cantidad de tráfico que generen.

8. En el modelo de representación del conocimiento basado en herencia se puede asegurar que:

a) La cantidad de padres que puede tener un nodo no es mayor que 4.

b) Las conclusiones pueden ser canceladas si el grafo es ambiguo.

c) Solo se usa en entornos referentes a la biología.

d) El razonamiento deducido está respaldado por al menos un camino dentro del grafo.

9. La Web 1.0 se caracteriza por:

a) Las actualizaciones del contenido de los sitios web se efectúa de forma manual dentro del código HTML

b) Los sitios web se centran en brindar información en lugar de facilitar la colaboración o participación de los usuarios.

c) Uso de comunidades virtuales para popularizar los sitios web de noticias.

d) Los propietarios de los sitios web proporcionan contenido de forma periódica.

10. En la RI en el contexto de Big Data, se puede asegurar que:

a) El almacenamiento distribuido centraliza todos los datos en un único servidor para facilitar su gestión y mantenimiento.

b) El almacenamiento distribuido ofrece ventajas significativas en términos de escalabilidad y rendimiento en comparación con el almacenamiento centralizado.

c) El almacenamiento distribuido reparte los datos en múltiples servidores para mejorar la disponibilidad y la redundancia del sistema.

d) El almacenamiento distribuido es una técnica obsoleta.

11. El propósito de la política de revisitado en los Web Crawlers es:

a) Determinar la frecuencia con la que los crawlers deben volver a visitar una URL específica para mantener la información actualizada en el índice de búsqueda.

b) Establecer reglas sobre el tiempo máximo que los crawlers pueden pasar en un sitio web durante cada visita para evitar sobrecargar los servidores.

c) Limitar el acceso de los crawlers a ciertos servidores luego de visitar las páginas alojadas en estos.

d) Definir la prioridad de rastreo de las URLs, determinando el orden en que los crawlers visitan y procesan cada página web.

12. Dentro del ecosistema de Hadoop, el HDFS se caracteriza por:

a) El modelo de acceso y de escritura de datos en tiempo real.

b) La tolerancia a fallos mediante la replicación de datos.

c) El almacenamiento exclusivo para archivos de texto.

d) La capacidad ilimitada de almacenamiento.

13. En una biblioteca digital se necesita implementar un sistema que permita a los usuarios encontrar libros y artículos científicos de forma rápida y precisa. Los documentos están en diversos formatos, incluyendo PDF, EPUB y HTML. Se requiere seleccionar un algoritmo de indexación adecuado para el sistema, por lo que el programador designado para la implementación debe considerar:

a) La velocidad de indexación y recuperación de datos.

b) La capacidad para manejar documentos en diferentes formatos.

c) La complejidad del algoritmo en términos de implementación y mantenimiento.

d) La capacidad del algoritmo para procesar imágenes incrustadas o referenciadas en los ficheros.

14. En el diseño de un videojuego de roles (RPG) en el que los personajes tienen atributos como salud, fuerza y velocidad, ¿cuál de las siguientes opciones representa mejor una implementación de la representación del conocimiento orientado a objetos?

a) Cada personaje se representa como una lista de cadenas de texto que describen sus características físicas y habilidades.

b) Cada personaje se representa como una función que calcula sus atributos en función de su nivel y experiencia.

c) Cada personaje se representa como una matriz de números que almacena sus valores de atributos.

d) Cada personaje se representa como un objeto con propiedades como salud, fuerza y velocidad, y métodos para modificar y consultar estos valores.

15. La Web 3 se conoce como:

a) Internet de las cosas.

b) Web de solo lectura.

c) Web semántica.

d) Web de escritura-lectura.

16. Considerando las dimensiones y desafíos inherentes a Big Data puede afirmarse, tomando en cuenta las características clave y las implicaciones para su procesamiento y análisis, que:
    a) Big Data se caracteriza principalmente por su pequeño volumen y uniformidad, permitiendo un procesamiento eficiente con mínimas adaptaciones de las herramientas de análisis de datos tradicionales.
    b) Big Data no desafía la capacidad de las herramientas tradicionales de procesamiento de datos para capturar, almacenar, gestionar y analizar efectivamente la información, dada la evolución constante de las capacidades computacionales y algoritmos de optimización.
    c) Aunque Big Data puede incluir datos estructurados, su naturaleza se expande al incorporar grandes cantidades de datos no estructurados y semiestructurados, lo que exige el uso de tecnologías especializadas en almacenamiento y procesamiento como Hadoop y sistemas de bases de datos NoSQL.
    d) Además de su complejidad y diversidad, Big Data introduce desafíos significativos en términos de veracidad y variabilidad al requerir métodos avanzados de limpieza y validación de datos para asegurar la integridad del análisis.
17. La centralidad de intermediación de un nodo indica:
    a) La cantidad de nodos vecinos directos.
    b) La resistencia del nodo a fallos.
    c) El número total de conexiones entrantes y salientes.
    d) La frecuencia con la que un nodo actúa como puente en el camino más corto entre otros dos nodos.
18. Una buena práctica de SEO para mejorar el posicionamiento de un sitio web en los motores de búsqueda es:
    a) Obtener enlaces de otros sitios web relevantes y de calidad que apunten al sitio.
    b) Copiar contenido directamente de otros sitios web populares para aumentar la cantidad de páginas indexadas.
    c) Utilizar técnicas de encubrimiento para mostrar contenido diferente a lo indexado por los motores de búsqueda.
    d) Llenar el contenido con palabras clave irrelevantes para aumentar la densidad de palabras clave.
19. Para la RI, el análisis de las redes puede:
    a) Ayudar a identificar grupos de interés.
    b) Indicar la importancia de una entidad en la transmisión de la información.
    c) Ayudar a comprender la conectividad y la accesibilidad entre las entidades.
    d) Revelar patrones de influencia dentro de una comunidad.
20. La Web 2.0 se caracteriza por:
    a) La existencia de sitios web dinámicos e interactivos que permiten a los usuarios participar, comentar e interactuar tanto con los creadores de contenido como con otros usuarios.
    b) El uso de distintas tecnologías para crear experiencias web más interactivas y con mayor capacidad de respuesta.
    c) Las actualizaciones del contenido de los sitios web se efectúa de forma manual dentro del código HTML
    d) Las plataformas ofrecen experiencias a la medida, permitiendo a los usuarios personalizar sus perfiles, recibir recomendaciones ajustadas al contenido y participar en filtrado colaborativo.

Trabajo de Control
Sistemas de Recuperación de Información
Fecha: 15 de julio de 2024
Temario #11

Nombre: _____ Grupo: _____

Seleccione en cada caso la(s) respuesta(s) correcta(s). Considere también no marcar ninguna opción.

1. ¿Qué algoritmos permiten obtener información de una red?
   a) Índices de centralidad.
   b) Detección de comunidades.
   c) Hypertext Induced Topic Selection (HITS).
   d) PageRank.

2. En el contexto de la RI en redes sociales se puede afirmar que:
   a) El análisis de redes sociales se centra exclusivamente en la cantidad de seguidores que tiene un usuario en particular para determinar su influencia en la red.
   b) El análisis de centralidad de intermediación se utiliza para identificar usuarios que son importantes en una red social debido a su posición como "puentes" entre diferentes grupos de usuarios.
   c) El análisis de sentimientos se utiliza para determinar la popularidad de una publicación en redes sociales sin tener en cuenta la opinión de los usuarios.
   d) El análisis de redes sociales no es útil para comprender la difusión de información en una red social específica, para ello se utiliza la medida de centralidad de vector propio.

3. En el contexto de la representación del conocimiento basada en herencia, ¿qué caracteriza a la herencia cancelable?
   a) La herencia es el resultado del razonamiento no transitivo.
   b) Las conclusiones no están determinadas y dependen del nodo de interés.
   c) Las propiedades heredadas siempre se mantienen y no pueden anularse.
   d) No existe ambigüedad en las conclusiones obtenidas.

4. ¿Qué estrategia utilizan los Web Crawlers para asegurar un rastreo eficiente y respetuoso de los recursos de los sitios web?
   a) Ignorar completamente el archivo `robots.txt` de los sitios web.
   b) Visitar y rastrear todos los enlaces de una página web simultáneamente.
   c) Extraer únicamente contenido multimedia para reducir la carga en los servidores web.
   d) Seguir las directrices del archivo `robots.txt` y aplicar un retraso entre las solicitudes.

5. Se puede afirmar que:
   a) El término "Big Data" se refiere exclusivamente al volumen de datos que una organización maneja, sin tener en cuenta la velocidad, la variedad y la veracidad de los datos.
   b) Los SRI pueden beneficiarse de MapReduce para mejorar la RI relevante.
   c) MapReduce es un modelo de procesamiento distribuido utilizado para trabajar con grandes volúmenes de datos.
   d) Uno de los desafíos en el procesamiento de Big Data es la capacidad de gestionar y analizar datos provenientes de diversas fuentes y en diferentes formatos de manera eficiente.

6. El propósito de la política de revisitado en los Web Crawlers es:
   a) Determinar la frecuencia con la que los crawlers deben volver a visitar una URL específica para mantener la información actualizada en el índice de búsqueda.
   b) Establecer reglas sobre el tiempo máximo que los crawlers pueden pasar en un sitio web durante cada visita para evitar sobrecargar los servidores.
   c) Limitar el acceso de los crawlers a ciertos servidores luego de visitar las páginas alojadas en estos.
   d) Definir la prioridad de rastreo de las URLs, determinando el orden en que los crawlers visitan y procesan cada página web.

7. Sobre WordNet y su aplicación en el procesamiento del lenguaje natural, se puede afirmar que:
   a) El diseño de WordNet facilita su integración en aplicaciones multilingües de NLP, aunque su desarrollo original se centró en el inglés.
   b) Aunque WordNet es una herramienta valiosa en el NLP, su estructura no incluye información sobre la frecuencia de uso de las palabras en el lenguaje natural.
   c) Los synsets facilitan la identificación de relaciones semánticas entre palabras, como la hiperonimia y la meronimia, enriqueciendo tareas de NLP.

d) WordNet proporciona una base para la desambiguación semántica al agrupar palabras con significados similares en synsets.

8. En el contexto de la RI en redes, se puede afirmar que:
   a) La detección de comunidades no es relevante para la RI en redes, ya que se centra únicamente en la estructura de la red sin considerar el contenido.
   b) La detección de comunidades en una red siempre produce resultados objetivos y consistentes independientemente del algoritmo utilizado.
   c) La detección de comunidades ayuda a identificar grupos de nodos altamente conectados entre sí, lo que puede ser útil para comprender la estructura y el contenido de la red.
   d) La detección de comunidades solo se aplica a redes pequeñas y simples, no a redes grandes y complejas.

9. Se tiene un grafo donde cada nodo es un personaje de cierto libro de cuentos y la existencia de las aristas está definida si dos personajes aparecen en el mismo cuento. Se puede asegurar que:
   a) La centralidad de intermediación identifica a los personajes que actúan como conectores entre personajes de cuentos distintos.
   b) La centralidad de grado es útil para identificar los personajes con más conexiones dentro de la red, lo que puede indicar su importancia en el libro.
   c) La centralidad de cercanía ofrece una relación entre la cantidad de vecinos de un nodo con respecto a la longitud máxima de un camino dentro del grafo partiendo del nodo en cuestión.
   d) La centralidad de cercanía indica el grado de conexión de cada personaje con el resto de los personajes de los cuentos del libro.

10. Considerando las prácticas éticas y legales en el Web Scraping, se puede asegurar que:
    a) Web Scraping requiere considerar las políticas de `robots.txt` del sitio web objetivo.
    b) Es importante revisar y respetar los términos de servicio del sitio web, así como las leyes aplicables de protección de datos y derechos de autor, antes de realizar Web Scraping.
    c) Web Scraping sobre datos personales sin consentimiento es generalmente aceptado si los datos se utilizan con fines de investigación.
    d) La extracción de datos mediante Web Scraping siempre es legal, independientemente de las leyes locales sobre derechos de autor y privacidad de datos definidos en los sitios web.

11. En el modelo de representación del conocimiento basado en herencia se puede asegurar que:
    a) La cantidad de padres que puede tener un nodo no es mayor que 4.
    b) Las conclusiones pueden ser canceladas si el grafo es ambiguo.
    c) Solo se usa en entornos referentes a la biología.
    d) El razonamiento deducido está respaldado por al menos un camino dentro del grafo.

12. Si una red cumple la propiedad de ser un grafo de mundo pequeño, entonces se conoce que:
    a) El número de componentes fuertemente conexas está relacionado con la cantidad de grafos K_n presentes.
    b) La longitud media del camino entre todo par de vértices es pequeña.
    c) El grafo es un anillo regular de grado 5.
    d) La red posee pocos vértices.

13. El algoritmo de PageRank puede describirse como un procedimiento utilizado para:
    a) Determinar la velocidad de carga de una página web en un navegador.
    b) Calcular la relevancia de una página web en función de la cantidad y calidad de los enlaces que apuntan hacia ella.
    c) Prevenir el spam y el contenido no deseado en las páginas web.
    d) Clasificar las páginas web en función de su edad y autoridad.

14. Un SRI es capaz de:
    a) Crear los índices asociados a los datos sin tener que analizar cada dato.
    b) Generar índices invertidos de manera óptima sin considerar el contexto.
    c) Reducir el tiempo de indexación de los datos si utiliza servidores distribuidos dentro de la red para que cada uno ejecute la indexación del mismo conjunto de datos.
    d) No necesitar de ningún almacenamiento externo para alojar los índices de los datos.

15. Se tiene una red de ingredientes donde cada uno representa un nodo y las aristas simbolizan que los ingredientes forman parte de una misma receta. Se busca mejorar la experiencia culinaria mediante la elaboración de combinaciones de ingredientes más interesantes y creativas, para lo cual se debe:
    a) No considerar la centralidad de grado de los ingredientes, ya que todas las combinaciones de ingredientes son igualmente válidas.
    b) Utilizar la centralidad de grado para identificar los ingredientes menos conectados en la red y tomarlos en cuenta para su inclusión en las combinaciones.

c) No tener en cuenta la centralidad de grado de los ingredientes para no darle mayor importancia a los ingredientes más comunes.

d) Utilizar la centralidad de grado para identificar los ingredientes más populares en la red y crear combinaciones que incluyan una variedad de ingredientes menos comunes.

16. En el contexto del procesamiento de grandes conjuntos de datos, se puede asegurar que:

a) MapReduce es un enfoque para procesar datos de forma secuencial en un solo servidor para evitar problemas de concurrencia.

b) MapReduce divide una tarea en múltiples pasos de map y reduce que se ejecutan de forma secuencial en diferentes servidores para mejorar el rendimiento y la escalabilidad.

c) MapReduce no es adecuado para el procesamiento de datos no estructurados.

d) MapReduce solo puede manejar pequeñas cantidades de datos y no escala bien a grandes conjuntos de datos.

17. La Web 1.0 se caracteriza por:

a) Las actualizaciones del contenido de los sitios web se efectúa de forma manual dentro del código HTML

b) Los sitios web se centran en brindar información en lugar de facilitar la colaboración o participación de los usuarios.

c) Uso de comunidades virtuales para popularizar los sitios web de noticias.

d) Los propietarios de los sitios web proporcionan contenido de forma periódica.

18. ¿Qué es un "Uniform Resource Locator (URL) Frontier" en el contexto de Web Crawling?

a) Un protocolo que define cómo se deben formatear las URLs para el crawling.

b) Una lista prioritaria de URLs que aún no han sido visitadas por el crawler.

c) Una técnica para filtrar URLs irrelevantes y mejorar la eficiencia del crawling.

d) Una base de datos que almacena URLs únicas identificadas como recursos en la web.

19. Si se tiene el conjunto de páginas interconectadas $\{A \to B, C; B \to C; C \to A\}$, entonces la página con valor más alto de PageRank es:

a) A y B.

b) A.

c) C.

d) B.

20. La web actual se enfrenta a problemas como:

a) Presencia de grandes volúmenes de datos estructurados.

b) Presencia de una alta calidad en los datos.

c) Presencia de datos volátiles y distribuidos.

d) Heterogeneidad en los datos.

Trabajo de Control
Sistemas de Recuperación de Información
Fecha: 15 de julio de 2024
Temario #12

Nombre: _____ Grupo: _____

Seleccione en cada caso la(s) respuesta(s) correcta(s). Considere también no marcar ninguna opción.

1. ¿Qué es un "Uniform Resource Locator (URL) Frontier" en el contexto de Web Crawling?
   a) Un protocolo que define cómo se deben formatear las URLs para el crawling.
   b) Una lista prioritaria de URLs que aún no han sido visitadas por el crawler.
   c) Una técnica para filtrar URLs irrelevantes y mejorar la eficiencia del crawling.
   d) Una base de datos que almacena URLs únicas identificadas como recursos en la web.

2. La Web 1.0 se caracteriza por:
   a) Las actualizaciones del contenido de los sitios web se efectúa de forma manual dentro del código HTML
   b) Los sitios web se centran en brindar información en lugar de facilitar la colaboración o participación de los usuarios.
   c) Uso de comunidades virtuales para popularizar los sitios web de noticias.
   d) Los propietarios de los sitios web proporcionan contenido de forma periódica.

3. Dentro del ecosistema de Hadoop, el HDFS se caracteriza por:
   a) El modelo de acceso y de escritura de datos en tiempo real.
   b) La tolerancia a fallos mediante la replicación de datos.
   c) El almacenamiento exclusivo para archivos de texto.
   d) La capacidad ilimitada de almacenamiento.

4. Dentro del análisis de redes, la centralidad de grado mide la importancia de un nodo basándose en:
   a) El grado del nodo.
   b) La cantidad de veces que aparece el nodo en el camino mínimo entre cualquier par de nodos.
   c) La cantidad de vecinos del nodo.
   d) El número de aristas que posee el nodo.

5. Si una red cumple la propiedad de ser un grafo de mundo pequeño, entonces se conoce que:
   a) El número de componentes fuertemente conexas está relacionado con la cantidad de grafos K_n presentes.
   b) La longitud media del camino entre todo par de vértices es pequeña.
   c) El grafo es un anillo regular de grado 5.
   d) La red posee pocos vértices.

6. La web actual se enfrenta a problemas como:
   a) Presencia de grandes volúmenes de datos estructurados.
   b) Presencia de una alta calidad en los datos.
   c) Presencia de datos volátiles y distribuidos.
   d) Heterogeneidad en los datos.

7. Si se tiene el conjunto de páginas interconectadas $\{A \to B, C; B \to C; C \to A\}$, entonces la página con valor más alto de PageRank es:
   a) A y B.
   b) A.
   c) C.
   d) B.

8. Analizar una red permite:
   a) Detectar posibles tendencias antes de que se conviertan en tendencia.
   b) Obtener predicciones exactas de eventos futuros en mercados financieros.
   c) Evaluar la calidad del contenido de un sitio web.
   d) Encontrar nodos "sensibles" o críticos para la red.

9. En el contexto de la RI en redes sociales se puede afirmar que:
   a) El análisis de redes sociales se centra exclusivamente en la cantidad de seguidores que tiene un usuario en particular para determinar su influencia en la red.
   b) El análisis de centralidad de intermediación se utiliza para identificar usuarios que son importantes en una red social debido a su posición como "puentes" entre diferentes grupos de usuarios.
   c) El análisis de sentimientos se utiliza para determinar la popularidad de una publicación en redes sociales sin tener en cuenta la opinión de los usuarios.
   d) El análisis de redes sociales no es útil para comprender la difusión de información en una red social específica, para ello se utiliza la medida de centralidad de vector propio.

10. En un SRI la indexación:
    a) Consiste en asociar un identificador único a cada dato almacenado en el sistema.
    b) Mejora la experiencia del usuario.
    c) Permite la organización y la categorización de la información.
    d) Facilita la RI relevante.

11. ¿Qué algoritmos permiten obtener información de una red?
    a) Índices de centralidad.
    b) Detección de comunidades.
    c) Hypertext Induced Topic Selection (HITS).
    d) PageRank.

12. El concepto de *rank sink* en el algoritmo de PageRank representa páginas web:
    a) Con distribuciones uniformes de la puntuación de PageRank.
    b) Que tienen una puntuación de PageRank más alta que otras debido a la manipulación de enlaces entrantes y salientes.
    c) Con una baja calidad de contenido y una cantidad insuficiente de enlaces salientes, lo que las hace menos relevantes en los resultados de búsqueda.
    d) Con un alto número de enlaces salientes que no reciben enlaces entrantes, lo que puede afectar negativamente su puntuación de PageRank.

13. Sobre WordNet y su aplicación en el procesamiento del lenguaje natural, se puede afirmar que:
    a) El diseño de WordNet facilita su integración en aplicaciones multilingües de NLP, aunque su desarrollo original se centró en el inglés.
    b) Aunque WordNet es una herramienta valiosa en el NLP, su estructura no incluye información sobre la frecuencia de uso de las palabras en el lenguaje natural.
    c) Los synsets facilitan la identificación de relaciones semánticas entre palabras, como la hiperonimia y la meronimia, enriqueciendo tareas de NLP.
    d) WordNet proporciona una base para la desambiguación semántica al agrupar palabras con significados similares en synsets.

14. Un investigador necesita recopilar datos de múltiples sitios web para un estudio académico, pero se enfrenta a varios desafíos al realizar el proceso de Web Scraping de manera ética y legal. ¿Cuál de las siguientes opciones describe mejor uno de los desafíos asociados al proceso de Web Scraping?
    a) La necesidad de comprender la estructura del sitio web y su código HTML para extraer los datos correctamente.
    b) La disponibilidad limitada de datos en línea que se pueden extraer utilizando técnicas de Web Scraping.
    c) La dificultad para encontrar herramientas de Web Scraping gratuitas y fiables.
    d) La necesidad de estar montado sobre un scrawler que cumpla con todas las políticas.

15. En la RI en el contexto de Big Data, se puede asegurar que:
    a) La indexación distribuida divide los datos en múltiples fragmentos que se almacenan en varios nodos para permitir búsquedas paralelas y mejorar la escalabilidad.
    b) La indexación distribuida no ofrece ventajas en términos de rendimiento y escalabilidad en comparación con la indexación centralizada.
    c) La indexación distribuida es una técnica obsoleta.
    d) La indexación distribuida almacena todos los datos en un solo servidor para facilitar su acceso y búsqueda.

16. El componente responsable de la gestión de recursos y planificación de tareas en Hadoop es:
    a) MapReduce.
    b) Hadoop Common.
    c) HDFS.
    d) YARN.

17. En un sistema donde el conocimiento está definido a partir de reglas se puede asegurar que:
    a) El nivel de especificidad de las reglas está limitado.
    b) La representación de conocimiento está basada en la lógica proposicional.
    c) El razonamiento se activa en cadena hacia delante.
    d) El orden en que se definen las reglas no altera el razonamiento del sistema.

18. Para transformar el contenido no estructurado de las páginas web en datos estructurados, el Web Scraping:
    a) Extrae información basada en patrones de HTML/CSS.
    b) Analiza los protocolos de red.
    c) Interpreta el código JavaScript en tiempo real.
    d) Convierte de forma automática imágenes a texto.

19. Un SRI es capaz de:
    a) Crear los índices asociados a los datos sin tener que analizar cada dato.
    b) Generar índices invertidos de manera óptima sin considerar el contexto.

c) Reducir el tiempo de indexación de los datos si utiliza servidores distribuidos dentro de la red para que cada uno ejecute la indexación del mismo conjunto de datos.

d) No necesitar de ningún almacenamiento externo para alojar los índices de los datos.

20. En el diseño de un videojuego de roles (RPG) en el que los personajes tienen atributos como salud, fuerza y velocidad, ¿cuál de las siguientes opciones representa mejor una implementación de la representación del conocimiento orientado a objetos?

a) Cada personaje se representa como una lista de cadenas de texto que describen sus características físicas y habilidades.

b) Cada personaje se representa como una función que calcula sus atributos en función de su nivel y experiencia.

c) Cada personaje se representa como una matriz de números que almacena sus valores de atributos.

d) Cada personaje se representa como un objeto con propiedades como salud, fuerza y velocidad, y métodos para modificar y consultar estos valores.

Trabajo de Control
Sistemas de Recuperación de Información
Fecha: 15 de julio de 2024
Temario #13

Nombre: _____ Grupo: _____

Seleccione en cada caso la(s) respuesta(s) correcta(s). Considere también no marcar ninguna opción.

1. Un SRI es capaz de:
   a) Crear los índices asociados a los datos sin tener que analizar cada dato.
   b) Generar índices invertidos de manera óptima sin considerar el contexto.
   c) Reducir el tiempo de indexación de los datos si utiliza servidores distribuidos dentro de la red para que cada uno ejecute la indexación del mismo conjunto de datos.
   d) No necesitar de ningún almacenamiento externo para alojar los índices de los datos.

2. Sobre WordNet y su aplicación en el procesamiento del lenguaje natural, se puede afirmar que:
   a) El diseño de WordNet facilita su integración en aplicaciones multilingües de NLP, aunque su desarrollo original se centró en el inglés.
   b) Aunque WordNet es una herramienta valiosa en el NLP, su estructura no incluye información sobre la frecuencia de uso de las palabras en el lenguaje natural.
   c) Los synsets facilitan la identificación de relaciones semánticas entre palabras, como la hiperonimia y la meronimia, enriqueciendo tareas de NLP.
   d) WordNet proporciona una base para la desambiguación semántica al agrupar palabras con significados similares en synsets.

3. Para contribuir positivamente al posicionamiento orgánico de un sitio web en los motores de búsqueda se puede:
   a) Crear contenido relevante y de alta calidad que satisfaga las necesidades de información de los usuarios.
   b) Mejorar la velocidad de carga del sitio web y asegurar que sea *responsive* y fácil de usar en dispositivos móviles.
   c) Incluir una densidad alta de palabras clave para asegurar que el sitio web aparezca en tantas búsquedas como sea posible.
   d) Obtener enlaces entrantes de otros sitios web de alta autoridad y relevancia temática.

4. Dentro del ecosistema de Hadoop, el HDFS se caracteriza por:
   a) El modelo de acceso y de escritura de datos en tiempo real.
   b) La tolerancia a fallos mediante la replicación de datos.
   c) El almacenamiento exclusivo para archivos de texto.
   d) La capacidad ilimitada de almacenamiento.

5. La política de ordenación de URLs en los Web Crawlers tiene como aspecto fundamental:
   a) Limitar el acceso de los crawlers a ciertas secciones de un sitio web, evitando el rastreo de URLs consideradas menos importantes o sensibles.
   b) Definir la estructura de la URL de destino, asegurando que estén ordenadas alfabéticamente para facilitar la navegación y la indexación.
   c) Establecer la prioridad de rastreo de las URLs, determinando el orden en que los crawlers visitan y procesan cada página web.
   d) Determinar la forma en que los crawlers asignan un valor de relevancia a cada URL para clasificarlas en el índice de búsqueda.

6. Al "relajar" el concepto de clique en la detección de comunidades se intenta solucionar:
   a) La imposibilidad de conectar nodos distantes.
   b) La necesidad de datos externos para analizar la red.
   c) La uniformidad de los nodos en términos de grado.
   d) El solapamiento y la complejidad computacional.

7. En un grafo una comunidad es:
   a) Un conjunto de nodos que no comparten ninguna similitud estructural o funcional entre sí.
   b) Un conjunto de nodos aislados.
   c) Un conjunto de nodos altamente conectados que forman un subgrafo completamente independiente del resto de la red.
   d) Un conjunto de nodos que están más densamente interconectados entre sí que con los nodos fuera del conjunto.

8. La computación evolutiva:
   a) No es aplicable en la RI debido a la complejidad de los algoritmos evolutivos.
   b) Solo puede manejar conjuntos de datos pequeños y no es escalable a grandes volúmenes de datos.

c) Utiliza algoritmos para buscar soluciones óptimas en grandes espacios de búsqueda, lo que la hace adecuada para problemas de optimización en la RI.

d) Es útil solo para problemas de clasificación de documentos y no para otras tareas de RI en general.

9. La premisa básica del algoritmo de PageRank para clasificar páginas web en los resultados de búsqueda es:

a) La antigüedad de la página web es el principal factor para determinar su clasificación.

b) El contenido y la relevancia de las palabras clave en la página web determinan su posición.

c) Los enlaces entrantes a una página web desde otras páginas contribuyen a su importancia y clasificación.

d) La cantidad de visitas que recibe una página web determina su clasificación.

10. En el modelo de representación del conocimiento basado en herencia se puede asegurar que:

a) La cantidad de padres que puede tener un nodo no es mayor que 4.

b) Las conclusiones pueden ser canceladas si el grafo es ambiguo.

c) Solo se usa en entornos referentes a la biología.

d) El razonamiento deducido está respaldado por al menos un camino dentro del grafo.

11. ¿Qué estrategia utilizan los Web Crawlers para asegurar un rastreo eficiente y respetuoso de los recursos de los sitios web?

a) Ignorar completamente el archivo `robots.txt` de los sitios web.

b) Visitar y rastrear todos los enlaces de una página web simultáneamente.

c) Extraer únicamente contenido multimedia para reducir la carga en los servidores web.

d) Seguir las directrices del archivo `robots.txt` y aplicar un retraso entre las solicitudes.

12. Para la RI, el análisis de las redes puede:

a) Ayudar a identificar grupos de interés.

b) Indicar la importancia de una entidad en la transmisión de la información.

c) Ayudar a comprender la conectividad y la accesibilidad entre las entidades.

d) Revelar patrones de influencia dentro de una comunidad.

13. En un sistema donde el conocimiento está orientado a objetos se puede asegurar que:

a) Se enfatiza la atención a la información de la cual se extrajo el conocimiento.

b) Los marcos y las bandas son las estructuras utilizadas para la representación del modelo.

c) No es posible definir especializaciones de los objetos de la vida real dentro del sistema.

d) Las funciones de agregación dificultan poder establecer relaciones entre los objetos.

14. Una buena práctica de SEO para mejorar el posicionamiento de un sitio web en los motores de búsqueda es:

a) Obtener enlaces de otros sitios web relevantes y de calidad que apunten al sitio.

b) Copiar contenido directamente de otros sitios web populares para aumentar la cantidad de páginas indexadas.

c) Utilizar técnicas de encubrimiento para mostrar contenido diferente a lo indexado por los motores de búsqueda.

d) Llenar el contenido con palabras clave irrelevantes para aumentar la densidad de palabras clave.

15. ¿Cuál de las siguientes opciones describe mejor la diferencia clave entre Web Crawling y Web Scraping?

a) Web Crawling se centra en la exploración y recopilación de enlaces de múltiples sitios web, mientras que Web Scraping se enfoca en la extracción específica de datos de páginas web individuales.

b) Web Crawling se realiza utilizando herramientas de automatización como Selenium WebDriver, mientras que Web Scraping se lleva a cabo mediante el análisis de HTML y CSS.

c) Web Scraping implica el análisis de la estructura y el contenido de las páginas web para extraer datos, mientras que Web Crawling se refiere a la descarga y almacenamiento de páginas web completas.

d) Web Scraping es más eficaz para rastrear e indexar contenido web para motores de búsqueda, mientras que Web Crawling se utiliza principalmente para la extracción de datos en proyectos de investigación.

16. El concepto de *rank sink* en el algoritmo de PageRank representa páginas web:

a) Con distribuciones uniformes de la puntuación de PageRank.

b) Que tienen una puntuación de PageRank más alta que otras debido a la manipulación de enlaces entrantes y salientes.

c) Con una baja calidad de contenido y una cantidad insuficiente de enlaces salientes, lo que las hace menos relevantes en los resultados de búsqueda.

d) Con un alto número de enlaces salientes que no reciben enlaces entrantes, lo que puede afectar negativamente su puntuación de PageRank.

17. En una empresa comercial, se tiene una red donde los nodos corresponden a los empleados y las aristas representan la frecuencia con la que colaboran en las ventas. Se busca mejorar la colaboración entre los empleados para aumentar las ventas totales, por lo que la directiva debe:

    a) Utilizar la centralidad de cercanía para identificar los empleados menos cercanos a otros en la red y asignarles tareas individuales para evitar posibles conflictos y desacuerdos en el proceso de colaboración.

    b) Utilizar la centralidad de cercanía para identificar los empleados más cercanos a otros en la red y promover la colaboración entre ellos, facilitando así la comunicación y el intercambio de conocimientos para mejorar las ventas.

    c) No considerar la centralidad de cercanía de los empleados al diseñar estrategias de colaboración, ya que todos los empleados tienen el mismo acceso a la información y los recursos.

    d) No tener en cuenta la centralidad de cercanía de los empleados al diseñar estrategias de colaboración, ya que esto podría introducir complicaciones adicionales en el proceso de trabajo.

18. Considerando las prácticas éticas y legales en el Web Scraping, se puede asegurar que:

    a) Web Scraping requiere considerar las políticas de `robots.txt` del sitio web objetivo.

    b) Es importante revisar y respetar los términos de servicio del sitio web, así como las leyes aplicables de protección de datos y derechos de autor, antes de realizar Web Scraping.

    c) Web Scraping sobre datos personales sin consentimiento es generalmente aceptado si los datos se utilizan con fines de investigación.

    d) La extracción de datos mediante Web Scraping siempre es legal, independientemente de las leyes locales sobre derechos de autor y privacidad de datos definidos en los sitios web.

19. Considerando las dimensiones y desafíos inherentes a Big Data puede afirmarse, tomando en cuenta las características clave y las implicaciones para su procesamiento y análisis, que:

    a) Big Data se caracteriza principalmente por su pequeño volumen y uniformidad, permitiendo un procesamiento eficiente con mínimas adaptaciones de las herramientas de análisis de datos tradicionales.

    b) Big Data no desafía la capacidad de las herramientas tradicionales de procesamiento de datos para capturar, almacenar, gestionar y analizar efectivamente la información, dada la evolución constante de las capacidades computacionales y algoritmos de optimización.

    c) Aunque Big Data puede incluir datos estructurados, su naturaleza se expande al incorporar grandes cantidades de datos no estructurados y semiestructurados, lo que exige el uso de tecnologías especializadas en almacenamiento y procesamiento como Hadoop y sistemas de bases de datos NoSQL.

    d) Además de su complejidad y diversidad, Big Data introduce desafíos significativos en términos de veracidad y variabilidad al requerir métodos avanzados de limpieza y validación de datos para asegurar la integridad del análisis.

20. Al realizar la optimización de contenido para SEO debe considerarse:

    a) Seleccionar las palabras clave al azar.

    b) Utilizar etiquetas de título y meta descripciones únicas y relevantes para cada página.

    c) Incluir palabras clave de manera excesiva en el contenido para mejorar el posicionamiento.

    d) Crear contenido valioso y original que satisfaga las necesidades de los usuarios.

Trabajo de Control
Sistemas de Recuperación de Información
Fecha: 15 de julio de 2024
Temario #14

Nombre: _____ Grupo: _____

Seleccione en cada caso la(s) respuesta(s) correcta(s). Considere también no marcar ninguna opción.

1. Si una red cumple la propiedad de ser un grafo de mundo pequeño, entonces se conoce que:
   a) El número de componentes fuertemente conexas está relacionado con la cantidad de grafos K_n presentes.
   b) La longitud media del camino entre todo par de vértices es pequeña.
   c) El grafo es un anillo regular de grado 5.
   d) La red posee pocos vértices.

2. Para transformar el contenido no estructurado de las páginas web en datos estructurados, el Web Scraping:
   a) Extrae información basada en patrones de HTML/CSS.
   b) Analiza los protocolos de red.
   c) Interpreta el código JavaScript en tiempo real.
   d) Convierte de forma automática imágenes a texto.

3. En una biblioteca digital se necesita implementar un sistema que permita a los usuarios encontrar libros y artículos científicos de forma rápida y precisa. Los documentos están en diversos formatos, incluyendo PDF, EPUB y HTML. Se requiere seleccionar un algoritmo de indexación adecuado para el sistema, por lo que el programador designado para la implementación debe considerar:
   a) La velocidad de indexación y recuperación de datos.
   b) La capacidad para manejar documentos en diferentes formatos.
   c) La complejidad del algoritmo en términos de implementación y mantenimiento.
   d) La capacidad del algoritmo para procesar imágenes incrustadas o referenciadas en los ficheros.

4. Al realizar la optimización de contenido para SEO debe considerarse:
   a) Seleccionar las palabras clave al azar.
   b) Utilizar etiquetas de título y meta descripciones únicas y relevantes para cada página.
   c) Incluir palabras clave de manera excesiva en el contenido para mejorar el posicionamiento.
   d) Crear contenido valioso y original que satisfaga las necesidades de los usuarios.

5. Para la RI, el análisis de las redes puede:
   a) Ayudar a identificar grupos de interés.
   b) Indicar la importancia de una entidad en la transmisión de la información.
   c) Ayudar a comprender la conectividad y la accesibilidad entre las entidades.
   d) Revelar patrones de influencia dentro de una comunidad.

6. La centralidad de intermediación de un nodo indica:
   a) La cantidad de nodos vecinos directos.
   b) La resistencia del nodo a fallos.
   c) El número total de conexiones entrantes y salientes.
   d) La frecuencia con la que un nodo actúa como puente en el camino más corto entre otros dos nodos.

7. La Web 3 se conoce como:
   a) Internet de las cosas.
   b) Web de solo lectura.
   c) Web semántica.
   d) Web de escritura-lectura.

8. Para contribuir positivamente al posicionamiento orgánico de un sitio web en los motores de búsqueda se puede:
   a) Crear contenido relevante y de alta calidad que satisfaga las necesidades de información de los usuarios.
   b) Mejorar la velocidad de carga del sitio web y asegurar que sea *responsive* y fácil de usar en dispositivos móviles.
   c) Incluir una densidad alta de palabras clave para asegurar que el sitio web aparezca en tantas búsquedas como sea posible.
   d) Obtener enlaces entrantes de otros sitios web de alta autoridad y relevancia temática.

9. La transición de la Web 1.0 a la Web 2.0 se caracterizó principalmente por:
   a) El aumento en la velocidad de conexión a internet, que permitió una mejor calidad de las páginas web.
   b) La disminución de la importancia de los motores de búsqueda en la navegación web.
   c) El cambio de páginas web estáticas a dinámicas, permitiendo la interacción del usuario y la generación de contenido.

d) La reducción en el uso de HTML y CSS en el desarrollo de sitios web.

10. En la RI en el contexto de Big Data, se puede asegurar que:
    a) El almacenamiento distribuido centraliza todos los datos en un único servidor para facilitar su gestión y mantenimiento.
    b) El almacenamiento distribuido ofrece ventajas significativas en términos de escalabilidad y rendimiento en comparación con el almacenamiento centralizado.
    c) El almacenamiento distribuido reparte los datos en múltiples servidores para mejorar la disponibilidad y la redundancia del sistema.
    d) El almacenamiento distribuido es una técnica obsoleta.

11. En una red de transporte donde cada nodo es una parada de autobús y las aristas representan si existe un carro que pasa por ambos sitios, ¿qué puede mejorar el sistema de transporte?
    a) No considerar la centralidad de intermediación de las estaciones de transporte público al planificar rutas y horarios, ya que no tiene impacto en las conexiones entre las paradas.
    b) Utilizar el grafo inducido de los nodos con mayor valor en la centralidad de grado para aplicar la centralidad de intermediación con el propósito de reforzar las paradas con mayor tráfico.
    c) Utilizar la centralidad de intermediación para identificar las paradas de transferencia clave y establecer nuevas conexiones entre diferentes líneas de transporte público.
    d) No implementar sistemas de información en tiempo real para los usuarios, ya que pueden aumentar la carga de los trabajadores.

12. La integración de Hadoop y MapReduce en la RI trae como ventaja:
    a) La eliminación de la necesidad de sistemas de bases de datos.
    b) La posibilidad de la extracción de información relevante y la generación de resultados significativos de grandes conjuntos de datos.
    c) La garantía de la privacidad absoluta de los datos procesados.
    d) La reducción de los costos operativos a cero.

13. En el modelo de representación del conocimiento basado en herencia se puede asegurar que:
    a) La cantidad de padres que puede tener un nodo no es mayor que 4.
    b) Las conclusiones pueden ser canceladas si el grafo es ambiguo.
    c) Solo se usa en entornos referentes a la biología.
    d) El razonamiento deducido está respaldado por al menos un camino dentro del grafo.

14. En el contexto de la RI en redes, se puede afirmar que:
    a) La detección de comunidades no es relevante para la RI en redes, ya que se centra únicamente en la estructura de la red sin considerar el contenido.
    b) La detección de comunidades en una red siempre produce resultados objetivos y consistentes independientemente del algoritmo utilizado.
    c) La detección de comunidades ayuda a identificar grupos de nodos altamente conectados entre sí, lo que puede ser útil para comprender la estructura y el contenido de la red.
    d) La detección de comunidades solo se aplica a redes pequeñas y simples, no a redes grandes y complejas.

15. Considerando las dimensiones y desafíos inherentes a Big Data puede afirmarse, tomando en cuenta las características clave y las implicaciones para su procesamiento y análisis, que:
    a) Big Data se caracteriza principalmente por su pequeño volumen y uniformidad, permitiendo un procesamiento eficiente con mínimas adaptaciones de las herramientas de análisis de datos tradicionales.
    b) Big Data no desafía la capacidad de las herramientas tradicionales de procesamiento de datos para capturar, almacenar, gestionar y analizar efectivamente la información, dada la evolución constante de las capacidades computacionales y algoritmos de optimización.
    c) Aunque Big Data puede incluir datos estructurados, su naturaleza se expande al incorporar grandes cantidades de datos no estructurados y semiestructurados, lo que exige el uso de tecnologías especializadas en almacenamiento y procesamiento como Hadoop y sistemas de bases de datos NoSQL.
    d) Además de su complejidad y diversidad, Big Data introduce desafíos significativos en términos de veracidad y variabilidad al requerir métodos avanzados de limpieza y validación de datos para asegurar la integridad del análisis.

16. La web actual se enfrenta a problemas como:
    a) Presencia de grandes volúmenes de datos estructurados.
    b) Presencia de una alta calidad en los datos.
    c) Presencia de datos volátiles y distribuidos.
    d) Heterogeneidad en los datos.

17. Una plataforma de comercio electrónico desea mejorar la experiencia del usuario al permitir una navegación más personalizada y contextualizada. Actualmente los usuarios tienen dificultades para encontrar productos relevantes debido a la gran cantidad de opciones disponibles. La empresa está interesada en implementar características de la Web 2.5 y la Web Semántica para abordar este problema. ¿Qué características podrían ayudar para ofrecer una navegación más personalizada y contextualizada?
    a) La implementación de ontologías y metadatos para enriquecer la descripción de productos y mejorar la precisión de las recomendaciones.
    b) La optimización de la velocidad de carga del sitio web para mejorar la experiencia del usuario y reducir el abandono del carrito de compra.
    c) La integración de redes sociales para permitir la recomendación de productos basada en las preferencias de amigos y contactos.
    d) La incorporación de los productos en tendencia en el mercado.

18. El algoritmo Hypertext Induced Topic Selection (HITS) intenta buscar nodos especiales. Estos son conocidos como:
    a) Autoridades.
    b) Hubs.
    c) Centrales.
    d) Sensibles.

19. El concepto de *rank sink* en el algoritmo de PageRank representa páginas web:
    a) Con distribuciones uniformes de la puntuación de PageRank.
    b) Que tienen una puntuación de PageRank más alta que otras debido a la manipulación de enlaces entrantes y salientes.
    c) Con una baja calidad de contenido y una cantidad insuficiente de enlaces salientes, lo que las hace menos relevantes en los resultados de búsqueda.
    d) Con un alto número de enlaces salientes que no reciben enlaces entrantes, lo que puede afectar negativamente su puntuación de PageRank.

20. La premisa básica del algoritmo de PageRank para clasificar páginas web en los resultados de búsqueda es:
    a) La antigüedad de la página web es el principal factor para determinar su clasificación.
    b) El contenido y la relevancia de las palabras clave en la página web determinan su posición.
    c) Los enlaces entrantes a una página web desde otras páginas contribuyen a su importancia y clasificación.
    d) La cantidad de visitas que recibe una página web determina su clasificación.

Trabajo de Control
Sistemas de Recuperación de Información
Fecha: 15 de julio de 2024
Temario #15

Nombre: _____ Grupo: _____

Seleccione en cada caso la(s) respuesta(s) correcta(s). Considere también no marcar ninguna opción.

1. Dentro del análisis de redes, la centralidad de grado mide la importancia de un nodo basándose en:
   a) El grado del nodo.
   b) La cantidad de veces que aparece el nodo en el camino mínimo entre cualquier par de nodos.
   c) La cantidad de vecinos del nodo.
   d) El número de aristas que posee el nodo.

2. En un sistema de control de tráfico urbano basado en reglas, ¿cuál de las siguientes reglas sería más efectiva para manejar situaciones de congestión en una intersección?
   a) Si hay pocos vehículos en la intersección, reducir el tiempo de los semáforos en verde.
   b) Si hay muchos vehículos en la intersección, aumentar el tiempo de los semáforos en verde.
   c) Si hay un vehículo de emergencia en la intersección, detener todos los demás vehículos.
   d) Si hay muchos peatones cruzando la intersección, reducir el tiempo de los semáforos en rojo.

3. No se considera como técnica para detectar comunidades en una red:
   a) Analizar la mutualidad de los enlaces.
   b) Usar el agrupamiento jerárquico.
   c) Utilizar el algoritmo de K-Means.
   d) Encontrar cliques de vértices de grado par.

4. En un SRI la indexación:
   a) Consiste en asociar un identificador único a cada dato almacenado en el sistema.
   b) Mejora la experiencia del usuario.
   c) Permite la organización y la categorización de la información.
   d) Facilita la RI relevante.

5. En la RI en el contexto de Big Data, se puede asegurar que:
   a) El almacenamiento distribuido centraliza todos los datos en un único servidor para facilitar su gestión y mantenimiento.
   b) El almacenamiento distribuido ofrece ventajas significativas en términos de escalabilidad y rendimiento en comparación con el almacenamiento centralizado.
   c) El almacenamiento distribuido reparte los datos en múltiples servidores para mejorar la disponibilidad y la redundancia del sistema.
   d) El almacenamiento distribuido es una técnica obsoleta.

6. Al "relajar" el concepto de clique en la detección de comunidades se intenta solucionar:
   a) La imposibilidad de conectar nodos distantes.
   b) La necesidad de datos externos para analizar la red.
   c) La uniformidad de los nodos en términos de grado.
   d) El solapamiento y la complejidad computacional.

7. El propósito de la política de revisitado en los Web Crawlers es:
   a) Determinar la frecuencia con la que los crawlers deben volver a visitar una URL específica para mantener la información actualizada en el índice de búsqueda.
   b) Establecer reglas sobre el tiempo máximo que los crawlers pueden pasar en un sitio web durante cada visita para evitar sobrecargar los servidores.
   c) Limitar el acceso de los crawlers a ciertos servidores luego de visitar las páginas alojadas en estos.
   d) Definir la prioridad de rastreo de las URLs, determinando el orden en que los crawlers visitan y procesan cada página web.

8. Se puede afirmar que:
   a) No es posible identificar subgrupos dentro de una red utilizando análisis de redes.
   b) La cantidad de conexiones de un nodo siempre indica su influencia en la red.
   c) El tamaño de una red es siempre indicativo de su efectividad en la transmisión de información.
   d) Todas las relaciones en una red tienen la misma importancia para el análisis.

9. En el diseño de un videojuego de roles (RPG) en el que los personajes tienen atributos como salud, fuerza y velocidad, ¿cuál de las siguientes opciones representa mejor una implementación de la representación del conocimiento orientado a objetos?

a) Cada personaje se representa como una lista de cadenas de texto que describen sus características físicas y habilidades.

b) Cada personaje se representa como una función que calcula sus atributos en función de su nivel y experiencia.

c) Cada personaje se representa como una matriz de números que almacena sus valores de atributos.

d) Cada personaje se representa como un objeto con propiedades como salud, fuerza y velocidad, y métodos para modificar y consultar estos valores.

10. Una plataforma de comercio electrónico desea mejorar la experiencia del usuario al permitir una navegación más personalizada y contextualizada. Actualmente los usuarios tienen dificultades para encontrar productos relevantes debido a la gran cantidad de opciones disponibles. La empresa está interesada en implementar características de la Web 2.5 y la Web Semántica para abordar este problema. ¿Qué características podrían ayudar para ofrecer una navegación más personalizada y contextualizada?

a) La implementación de ontologías y metadatos para enriquecer la descripción de productos y mejorar la precisión de las recomendaciones.

b) La optimización de la velocidad de carga del sitio web para mejorar la experiencia del usuario y reducir el abandono del carrito de compra.

c) La integración de redes sociales para permitir la recomendación de productos basada en las preferencias de amigos y contactos.

d) La incorporación de los productos en tendencia en el mercado.

11. Con respecto a WordNet y su contribución a las ontologías y la representación del conocimiento en los sistemas de RI, puede afirmarse que:

a) En WordNet, una palabra está asociada a un conjunto de sinónimos (synsets), siendo estas palabras intercambiables en un contexto.

b) La integración de WordNet en sistemas de RI permite la expansión de consultas y la mejora de la precisión de los resultados al entender mejor el significado de los términos de búsqueda a través de su contexto semántico y las relaciones entre palabras.

c) WordNet diferencia claramente entre los significados de palabras según su uso en diferentes contextos, lo que permite aplicaciones avanzadas en desambiguación semántica más allá de los sistemas de recomendación y RI.

d) Una limitación de WordNet en la representación del conocimiento es su enfoque en el idioma inglés, lo que plantea desafíos en la aplicación global y la interoperabilidad con sistemas de información multilingües.

12. Referente al Web Crawler puede afirmarse que:

a) Los hipervínculos encontrados en cada sitio web que no pertenecen al dominio donde fueron encontrados se desechan, puesto que no expande el conjunto de URLs sin visitar.

b) Las páginas visitadas no se procesan nunca más.

c) No necesita de un conjunto inicial de URLs para recorrer la Web.

d) No tiene como objetivo indexar y recopilar información de diferentes sitios web.

13. Una buena práctica de SEO para mejorar el posicionamiento de un sitio web en los motores de búsqueda es:

a) Obtener enlaces de otros sitios web relevantes y de calidad que apunten al sitio.

b) Copiar contenido directamente de otros sitios web populares para aumentar la cantidad de páginas indexadas.

c) Utilizar técnicas de encubrimiento para mostrar contenido diferente a lo indexado por los motores de búsqueda.

d) Llenar el contenido con palabras clave irrelevantes para aumentar la densidad de palabras clave.

14. El algoritmo Hypertext Induced Topic Selection (HITS) intenta buscar nodos especiales. Estos son conocidos como:

a) Autoridades.

b) Hubs.

c) Centrales.

d) Sensibles.

15. En la RI en el contexto de Big Data, se puede asegurar que:

a) La indexación distribuida divide los datos en múltiples fragmentos que se almacenan en varios nodos para permitir búsquedas paralelas y mejorar la escalabilidad.

b) La indexación distribuida no ofrece ventajas en términos de rendimiento y escalabilidad en comparación con la indexación centralizada.

c) La indexación distribuida es una técnica obsoleta.

d) La indexación distribuida almacena todos los datos en un solo servidor para facilitar su acceso y búsqueda.

16. Para transformar el contenido no estructurado de las páginas web en datos estructurados, el Web Scraping:

a) Extrae información basada en patrones de HTML/CSS.

b) Analiza los protocolos de red.

c) Interpreta el código JavaScript en tiempo real.

d) Convierte de forma automática imágenes a texto.

17. En una biblioteca digital se necesita implementar un sistema que permita a los usuarios encontrar libros y artículos científicos de forma rápida y precisa. Los documentos están en diversos formatos, incluyendo PDF, EPUB y HTML. Se requiere seleccionar un algoritmo de indexación adecuado para el sistema, por lo que el programador designado para la implementación debe considerar:

    a) La velocidad de indexación y recuperación de datos.

    b) La capacidad para manejar documentos en diferentes formatos.

    c) La complejidad del algoritmo en términos de implementación y mantenimiento.

    d) La capacidad del algoritmo para procesar imágenes incrustadas o referenciadas en los ficheros.

18. La Web 2.0 se caracteriza por:

    a) La existencia de sitios web dinámicos e interactivos que permiten a los usuarios participar, comentar e interactuar tanto con los creadores de contenido como con otros usuarios.

    b) El uso de distintas tecnologías para crear experiencias web más interactivas y con mayor capacidad de respuesta.

    c) Las actualizaciones del contenido de los sitios web se efectúa de forma manual dentro del código HTML

    d) Las plataformas ofrecen experiencias a la medida, permitiendo a los usuarios personalizar sus perfiles, recibir recomendaciones ajustadas al contenido y participar en filtrado colaborativo.

19. Implementar índices invertidos en un SRI asegura:

    a) Acelerar el proceso de búsqueda al permitir búsquedas directas por contenido en lugar de por título.

    b) Reducir la cantidad de espacio de almacenamiento necesario al comprimir los datos de los documentos.

    c) Facilitar la búsqueda sobre los datos que contienen términos específicos al mantener una lista de datos para cada término único.

    d) Incrementar la seguridad de los datos almacenados al dificultar el acceso directo a la información sin el índice correcto.

20. ¿Qué algoritmos permiten obtener información de una red?

    a) Índices de centralidad.

    b) Detección de comunidades.

    c) Hypertext Induced Topic Selection (HITS).

    d) PageRank.

Trabajo de Control
Sistemas de Recuperación de Información
Fecha: 15 de julio de 2024
Temario #16

Nombre: _____ Grupo: _____

Seleccione en cada caso la(s) respuesta(s) correcta(s). Considere también no marcar ninguna opción.

1. Los algoritmos para detectar comunidades en una red intentan:
   a) Buscar conjuntos donde cada nodo de un mismo conjunto tenga características similares al resto de los nodos del conjunto.
   b) Seleccionar aleatoriamente nodos de alto grado y sus vecinos.
   c) Buscar subgrafos tal que no incluyan nodos cuya ausencia desconecte al subgrafo.
   d) Encontrar grupos donde los nodos pertenecientes a los mismos grupos son cercanos bajo cierta métrica y lejanos con respecto a los nodos de otros grupos.

2. Para la RI, el análisis de las redes puede:
   a) Ayudar a identificar grupos de interés.
   b) Indicar la importancia de una entidad en la transmisión de la información.
   c) Ayudar a comprender la conectividad y la accesibilidad entre las entidades.
   d) Revelar patrones de influencia dentro de una comunidad.

3. Un investigador necesita recopilar datos de múltiples sitios web para un estudio académico, pero se enfrenta a varios desafíos al realizar el proceso de Web Scraping de manera ética y legal. ¿Cuál de las siguientes opciones describe mejor uno de los desafíos asociados al proceso de Web Scraping?
   a) La necesidad de comprender la estructura del sitio web y su código HTML para extraer los datos correctamente.
   b) La disponibilidad limitada de datos en línea que se pueden extraer utilizando técnicas de Web Scraping.
   c) La dificultad para encontrar herramientas de Web Scraping gratuitas y fiables.
   d) La necesidad de estar montado sobre un scrawler que cumpla con todas las políticas.

4. Dentro del análisis de redes, la centralidad de grado mide la importancia de un nodo basándose en:
   a) El grado del nodo.
   b) La cantidad de veces que aparece el nodo en el camino mínimo entre cualquier par de nodos.
   c) La cantidad de vecinos del nodo.
   d) El número de aristas que posee el nodo.

5. La Web 3 se conoce como:
   a) Internet de las cosas.
   b) Web de solo lectura.
   c) Web semántica.
   d) Web de escritura-lectura.

6. La computación evolutiva:
   a) No es aplicable en la RI debido a la complejidad de los algoritmos evolutivos.
   b) Solo puede manejar conjuntos de datos pequeños y no es escalable a grandes volúmenes de datos.
   c) Utiliza algoritmos para buscar soluciones óptimas en grandes espacios de búsqueda, lo que la hace adecuada para problemas de optimización en la RI.
   d) Es útil solo para problemas de clasificación de documentos y no para otras tareas de RI en general.

7. Con respecto a WordNet y su contribución a las ontologías y la representación del conocimiento en los sistemas de RI, puede afirmarse que:
   a) En WordNet, una palabra está asociada a un conjunto de sinónimos (synsets), siendo estas palabras intercambiables en un contexto.
   b) La integración de WordNet en sistemas de RI permite la expansión de consultas y la mejora de la precisión de los resultados al entender mejor el significado de los términos de búsqueda a través de su contexto semántico y las relaciones entre palabras.
   c) WordNet diferencia claramente entre los significados de palabras según su uso en diferentes contextos, lo que permite aplicaciones avanzadas en desambiguación semántica más allá de los sistemas de recomendación y RI.
   d) Una limitación de WordNet en la representación del conocimiento es su enfoque en el idioma inglés, lo que plantea desafíos en la aplicación global y la interoperabilidad con sistemas de información multilingües.

8. En el contexto de la representación del conocimiento basada en herencia, ¿qué caracteriza a la herencia cancelable?
   a) La herencia es el resultado del razonamiento no transitivo.

b) Las conclusiones no están determinadas y dependen del nodo de interés.

c) Las propiedades heredadas siempre se mantienen y no pueden anularse.

d) No existe ambigüedad en las conclusiones obtenidas.

9. No se considera como técnica para detectar comunidades en una red:

a) Analizar la mutualidad de los enlaces.

b) Usar el agrupamiento jerárquico.

c) Utilizar el algoritmo de K-Means.

d) Encontrar cliques de vértices de grado par.

10. En un sistema donde el conocimiento está orientado a objetos se puede asegurar que:

a) Se enfatiza la atención a la información de la cual se extrajo el conocimiento.

b) Los marcos y las bandas son las estructuras utilizadas para la representación del modelo.

c) No es posible definir especializaciones de los objetos de la vida real dentro del sistema.

d) Las funciones de agregación dificultan poder establecer relaciones entre los objetos.

11. En un SRI la indexación:

a) Consiste en asociar un identificador único a cada dato almacenado en el sistema.

b) Mejora la experiencia del usuario.

c) Permite la organización y la categorización de la información.

d) Facilita la RI relevante.

12. Si se tiene el conjunto de páginas interconectadas $\{A \to B, C; B \to C; C \to A\}$, entonces la página con valor más alto de PageRank es:

a) A y B.

b) A.

c) C.

d) B.

13. Analizar una red permite:

a) Detectar posibles tendencias antes de que se conviertan en tendencia.

b) Obtener predicciones exactas de eventos futuros en mercados financieros.

c) Evaluar la calidad del contenido de un sitio web.

d) Encontrar nodos "sensibles" o críticos para la red.

14. La transición de la Web 1.0 a la Web 2.0 se caracterizó principalmente por:

a) El aumento en la velocidad de conexión a internet, que permitió una mejor calidad de las páginas web.

b) La disminución de la importancia de los motores de búsqueda en la navegación web.

c) El cambio de páginas web estáticas a dinámicas, permitiendo la interacción del usuario y la generación de contenido.

d) La reducción en el uso de HTML y CSS en el desarrollo de sitios web.

15. Dentro del ecosistema de Hadoop, el HDFS se caracteriza por:

a) El modelo de acceso y de escritura de datos en tiempo real.

b) La tolerancia a fallos mediante la replicación de datos.

c) El almacenamiento exclusivo para archivos de texto.

d) La capacidad ilimitada de almacenamiento.

16. La web actual se enfrenta a problemas como:

a) Presencia de grandes volúmenes de datos estructurados.

b) Presencia de una alta calidad en los datos.

c) Presencia de datos volátiles y distribuidos.

d) Heterogeneidad en los datos.

17. El concepto de *rank sink* en el algoritmo de PageRank representa páginas web:

a) Con distribuciones uniformes de la puntuación de PageRank.

b) Que tienen una puntuación de PageRank más alta que otras debido a la manipulación de enlaces entrantes y salientes.

c) Con una baja calidad de contenido y una cantidad insuficiente de enlaces salientes, lo que las hace menos relevantes en los resultados de búsqueda.

d) Con un alto número de enlaces salientes que no reciben enlaces entrantes, lo que puede afectar negativamente su puntuación de PageRank.

18. El algoritmo Hypertext Induced Topic Selection (HITS) intenta buscar nodos especiales. Estos son conocidos como:

a) Autoridades.

b) Hubs.

c) Centrales.

d) Sensibles.

19. La afirmación que mejor describe la política de amabilidad en los Web Crawlers es:

a) Los crawlers se diseñan para acceder a sitios web sin restricciones y extraer datos de manera agresiva para su indexación.

b) La política de amabilidad de los Web Crawlers dicta que los crawlers deben priorizar ciertos tipos de contenido sobre otros, ignorando completamente ciertas páginas web.

c) La política de amabilidad establece pautas y reglas sobre cómo los crawlers deben interactuar con los sitios web para minimizar la carga del servidor y respetar las directivas de los administradores del sitio.

d) Los Web Crawlers son libres de recopilar datos de cualquier sitio web sin restricciones, independientemente de la cantidad de tráfico que generen.

20. ¿Qué algoritmos permiten obtener información de una red?

a) Índices de centralidad.

b) Detección de comunidades.

c) Hypertext Induced Topic Selection (HITS).

d) PageRank.

Trabajo de Control
Sistemas de Recuperación de Información
Fecha: 15 de julio de 2024
Temario #17

Nombre: _____ Grupo: _____

Seleccione en cada caso la(s) respuesta(s) correcta(s). Considere también no marcar ninguna opción.

1. No se considera como técnica para detectar comunidades en una red:
   a) Analizar la mutualidad de los enlaces.
   b) Usar el agrupamiento jerárquico.
   c) Utilizar el algoritmo de K-Means.
   d) Encontrar cliques de vértices de grado par.

2. Se tiene un grafo donde cada nodo es un personaje de cierto libro de cuentos y la existencia de las aristas está definida si dos personajes aparecen en el mismo cuento. Se puede asegurar que:
   a) La centralidad de intermediación identifica a los personajes que actúan como conectores entre personajes de cuentos distintos.
   b) La centralidad de grado es útil para identificar los personajes con más conexiones dentro de la red, lo que puede indicar su importancia en el libro.
   c) La centralidad de cercanía ofrece una relación entre la cantidad de vecinos de un nodo con respecto a la longitud máxima de un camino dentro del grafo partiendo del nodo en cuestión.
   d) La centralidad de cercanía indica el grado de conexión de cada personaje con el resto de los personajes de los cuentos del libro.

3. ¿Qué estrategia utilizan los Web Crawlers para asegurar un rastreo eficiente y respetuoso de los recursos de los sitios web?
   a) Ignorar completamente el archivo `robots.txt` de los sitios web.
   b) Visitar y rastrear todos los enlaces de una página web simultáneamente.
   c) Extraer únicamente contenido multimedia para reducir la carga en los servidores web.
   d) Seguir las directrices del archivo `robots.txt` y aplicar un retraso entre las solicitudes.

4. Se puede afirmar que:
   a) El término "Big Data" se refiere exclusivamente al volumen de datos que una organización maneja, sin tener en cuenta la velocidad, la variedad y la veracidad de los datos.
   b) Los SRI pueden beneficiarse de MapReduce para mejorar la RI relevante.
   c) MapReduce es un modelo de procesamiento distribuido utilizado para trabajar con grandes volúmenes de datos.
   d) Uno de los desafíos en el procesamiento de Big Data es la capacidad de gestionar y analizar datos provenientes de diversas fuentes y en diferentes formatos de manera eficiente.

5. En una empresa comercial, se tiene una red donde los nodos corresponden a los empleados y las aristas representan la frecuencia con la que colaboran en las ventas. Se busca mejorar la colaboración entre los empleados para aumentar las ventas totales, por lo que la directiva debe:
   a) Utilizar la centralidad de cercanía para identificar los empleados menos cercanos a otros en la red y asignarles tareas individuales para evitar posibles conflictos y desacuerdos en el proceso de colaboración.
   b) Utilizar la centralidad de cercanía para identificar los empleados más cercanos a otros en la red y promover la colaboración entre ellos, facilitando así la comunicación y el intercambio de conocimientos para mejorar las ventas.
   c) No considerar la centralidad de cercanía de los empleados al diseñar estrategias de colaboración, ya que todos los empleados tienen el mismo acceso a la información y los recursos.
   d) No tener en cuenta la centralidad de cercanía de los empleados al diseñar estrategias de colaboración, ya que esto podría introducir complicaciones adicionales en el proceso de trabajo.

6. En el contexto de la representación del conocimiento basada en herencia, ¿qué caracteriza a la herencia cancelable?
   a) La herencia es el resultado del razonamiento no transitivo.
   b) Las conclusiones no están determinadas y dependen del nodo de interés.
   c) Las propiedades heredadas siempre se mantienen y no pueden anularse.
   d) No existe ambigüedad en las conclusiones obtenidas.

7. En un sistema donde el conocimiento está definido a partir de reglas se puede asegurar que:

a) El nivel de especificidad de las reglas está limitado.

b) La representación de conocimiento está basada en la lógica proposicional.

c) El razonamiento se activa en cadena hacia delante.

d) El orden en que se definen las reglas no altera el razonamiento del sistema.

8. En un sistema donde el conocimiento está orientado a objetos se puede asegurar que:

a) Se enfatiza la atención a la información de la cual se extrajo el conocimiento.

b) Los marcos y las bandas son las estructuras utilizadas para la representación del modelo.

c) No es posible definir especializaciones de los objetos de la vida real dentro del sistema.

d) Las funciones de agregación dificultan poder establecer relaciones entre los objetos.

9. ¿Qué algoritmos permiten obtener información de una red?

a) Índices de centralidad.

b) Detección de comunidades.

c) Hypertext Induced Topic Selection (HITS).

d) PageRank.

10. El algoritmo de PageRank puede describirse como un procedimiento utilizado para:

a) Determinar la velocidad de carga de una página web en un navegador.

b) Calcular la relevancia de una página web en función de la cantidad y calidad de los enlaces que apuntan hacia ella.

c) Prevenir el spam y el contenido no deseado en las páginas web.

d) Clasificar las páginas web en función de su edad y autoridad.

11. Para contribuir positivamente al posicionamiento orgánico de un sitio web en los motores de búsqueda se puede:

a) Crear contenido relevante y de alta calidad que satisfaga las necesidades de información de los usuarios.

b) Mejorar la velocidad de carga del sitio web y asegurar que sea *responsive* y fácil de usar en dispositivos móviles.

c) Incluir una densidad alta de palabras clave para asegurar que el sitio web aparezca en tantas búsquedas como sea posible.

d) Obtener enlaces entrantes de otros sitios web de alta autoridad y relevancia temática.

12. La Web 1.0 se caracteriza por:

a) Las actualizaciones del contenido de los sitios web se efectúa de forma manual dentro del código HTML

b) Los sitios web se centran en brindar información en lugar de facilitar la colaboración o participación de los usuarios.

c) Uso de comunidades virtuales para popularizar los sitios web de noticias.

d) Los propietarios de los sitios web proporcionan contenido de forma periódica.

13. Sobre WordNet y su aplicación en el procesamiento del lenguaje natural, se puede afirmar que:

a) El diseño de WordNet facilita su integración en aplicaciones multilingües de NLP, aunque su desarrollo original se centró en el inglés.

b) Aunque WordNet es una herramienta valiosa en el NLP, su estructura no incluye información sobre la frecuencia de uso de las palabras en el lenguaje natural.

c) Los synsets facilitan la identificación de relaciones semánticas entre palabras, como la hiperonimia y la meronimia, enriqueciendo tareas de NLP.

d) WordNet proporciona una base para la desambiguación semántica al agrupar palabras con significados similares en synsets.

14. La integración de Hadoop y MapReduce en la RI trae como ventaja:

a) La eliminación de la necesidad de sistemas de bases de datos.

b) La posibilidad de la extracción de información relevante y la generación de resultados significativos de grandes conjuntos de datos.

c) La garantía de la privacidad absoluta de los datos procesados.

d) La reducción de los costos operativos a cero.

15. La política de ordenación de URLs en los Web Crawlers tiene como aspecto fundamental:

a) Limitar el acceso de los crawlers a ciertas secciones de un sitio web, evitando el rastreo de URLs consideradas menos importantes o sensibles.

b) Definir la estructura de la URL de destino, asegurando que estén ordenadas alfabéticamente para facilitar la navegación y la indexación.

c) Establecer la prioridad de rastreo de las URLs, determinando el orden en que los crawlers visitan y procesan cada página web.

d) Determinar la forma en que los crawlers asignan un valor de relevancia a cada URL para clasificarlas en el índice de búsqueda.

16. Con respecto a WordNet y su contribución a las ontologías y la representación del conocimiento en los sistemas de RI, puede afirmarse que:

a) En WordNet, una palabra está asociada a un conjunto de sinónimos (synsets), siendo estas palabras intercambiables en un contexto.

b) La integración de WordNet en sistemas de RI permite la expansión de consultas y la mejora de la precisión de los resultados al entender mejor el significado de los términos de búsqueda a través de su contexto semántico y las relaciones entre palabras.

c) WordNet diferencia claramente entre los significados de palabras según su uso en diferentes contextos, lo que permite aplicaciones avanzadas en desambiguación semántica más allá de los sistemas de recomendación y RI.

d) Una limitación de WordNet en la representación del conocimiento es su enfoque en el idioma inglés, lo que plantea desafíos en la aplicación global y la interoperabilidad con sistemas de información multilingües.

17. Sobre el algoritmo de PageRank, visto en clase, se puede afirmar que:

a) Evalúa la importancia de un sitio web en función de la calidad y cantidad de enlaces entrantes que recibe de otros sitios web.

b) Solo tiene en cuenta el contenido en un sitio web para determinar su relevancia en los resultados de búsqueda.

c) Asigna una puntuación alta a los sitios web que tienen un gran número de enlaces entrantes sin tener en cuenta la calidad de esos enlaces.

d) Asigna una puntuación baja a los sitios web que contienen muchos enlaces salientes, ya que indica una falta de relevancia.

18. Un investigador necesita recopilar datos de múltiples sitios web para un estudio académico, pero se enfrenta a varios desafíos al realizar el proceso de Web Scraping de manera ética y legal. ¿Cuál de las siguientes opciones describe mejor uno de los desafíos asociados al proceso de Web Scraping?

a) La necesidad de comprender la estructura del sitio web y su código HTML para extraer los datos correctamente.

b) La disponibilidad limitada de datos en línea que se pueden extraer utilizando técnicas de Web Scraping.

c) La dificultad para encontrar herramientas de Web Scraping gratuitas y fiables.

d) La necesidad de estar montado sobre un scrawler que cumpla con todas las políticas.

19. En una biblioteca digital se necesita implementar un sistema que permita a los usuarios encontrar libros y artículos científicos de forma rápida y precisa. Los documentos están en diversos formatos, incluyendo PDF, EPUB y HTML. Se requiere seleccionar un algoritmo de indexación adecuado para el sistema, por lo que el programador designado para la implementación debe considerar:

a) La velocidad de indexación y recuperación de datos.

b) La capacidad para manejar documentos en diferentes formatos.

c) La complejidad del algoritmo en términos de implementación y mantenimiento.

d) La capacidad del algoritmo para procesar imágenes incrustadas o referenciadas en los ficheros.

20. En la RI en el contexto de Big Data, se puede asegurar que:

a) El almacenamiento distribuido centraliza todos los datos en un único servidor para facilitar su gestión y mantenimiento.

b) El almacenamiento distribuido ofrece ventajas significativas en términos de escalabilidad y rendimiento en comparación con el almacenamiento centralizado.

c) El almacenamiento distribuido reparte los datos en múltiples servidores para mejorar la disponibilidad y la redundancia del sistema.

d) El almacenamiento distribuido es una técnica obsoleta.

Trabajo de Control
Sistemas de Recuperación de Información
Fecha: 15 de julio de 2024
Temario #18

Nombre: _____ Grupo: _____

Seleccione en cada caso la(s) respuesta(s) correcta(s). Considere también no marcar ninguna opción.

1. La centralidad de intermediación de un nodo indica:
   a) La cantidad de nodos vecinos directos.
   b) La resistencia del nodo a fallos.
   c) El número total de conexiones entrantes y salientes.
   d) La frecuencia con la que un nodo actúa como puente en el camino más corto entre otros dos nodos.

2. Se puede afirmar que:
   a) El término "Big Data" se refiere exclusivamente al volumen de datos que una organización maneja, sin tener en cuenta la velocidad, la variedad y la veracidad de los datos.
   b) Los SRI pueden beneficiarse de MapReduce para mejorar la RI relevante.
   c) MapReduce es un modelo de procesamiento distribuido utilizado para trabajar con grandes volúmenes de datos.
   d) Uno de los desafíos en el procesamiento de Big Data es la capacidad de gestionar y analizar datos provenientes de diversas fuentes y en diferentes formatos de manera eficiente.

3. En la RI en el contexto de Big Data, se puede asegurar que:
   a) El almacenamiento distribuido centraliza todos los datos en un único servidor para facilitar su gestión y mantenimiento.
   b) El almacenamiento distribuido ofrece ventajas significativas en términos de escalabilidad y rendimiento en comparación con el almacenamiento centralizado.
   c) El almacenamiento distribuido reparte los datos en múltiples servidores para mejorar la disponibilidad y la redundancia del sistema.
   d) El almacenamiento distribuido es una técnica obsoleta.

4. ¿Qué es un "Uniform Resource Locator (URL) Frontier" en el contexto de Web Crawling?
   a) Un protocolo que define cómo se deben formatear las URLs para el crawling.
   b) Una lista prioritaria de URLs que aún no han sido visitadas por el crawler.
   c) Una técnica para filtrar URLs irrelevantes y mejorar la eficiencia del crawling.
   d) Una base de datos que almacena URLs únicas identificadas como recursos en la web.

5. Con respecto a WordNet y su contribución a las ontologías y la representación del conocimiento en los sistemas de RI, puede afirmarse que:
   a) En WordNet, una palabra está asociada a un conjunto de sinónimos (synsets), siendo estas palabras intercambiables en un contexto.
   b) La integración de WordNet en sistemas de RI permite la expansión de consultas y la mejora de la precisión de los resultados al entender mejor el significado de los términos de búsqueda a través de su contexto semántico y las relaciones entre palabras.
   c) WordNet diferencia claramente entre los significados de palabras según su uso en diferentes contextos, lo que permite aplicaciones avanzadas en desambiguación semántica más allá de los sistemas de recomendación y RI.
   d) Una limitación de WordNet en la representación del conocimiento es su enfoque en el idioma inglés, lo que plantea desafíos en la aplicación global y la interoperabilidad con sistemas de información multilingües.

6. La computación evolutiva:
   a) No es aplicable en la RI debido a la complejidad de los algoritmos evolutivos.
   b) Solo puede manejar conjuntos de datos pequeños y no es escalable a grandes volúmenes de datos.
   c) Utiliza algoritmos para buscar soluciones óptimas en grandes espacios de búsqueda, lo que la hace adecuada para problemas de optimización en la RI.
   d) Es útil solo para problemas de clasificación de documentos y no para otras tareas de RI en general.

7. El propósito de la política de revisitado en los Web Crawlers es:
   a) Determinar la frecuencia con la que los crawlers deben volver a visitar una URL específica para mantener la información actualizada en el índice de búsqueda.
   b) Establecer reglas sobre el tiempo máximo que los crawlers pueden pasar en un sitio web durante cada visita para evitar sobrecargar los servidores.

c) Limitar el acceso de los crawlers a ciertos servidores luego de visitar las páginas alojadas en estos.

d) Definir la prioridad de rastreo de las URLs, determinando el orden en que los crawlers visitan y procesan cada página web.

8. La integración de Hadoop y MapReduce en la RI trae como ventaja:
   a) La eliminación de la necesidad de sistemas de bases de datos.
   b) La posibilidad de la extracción de información relevante y la generación de resultados significativos de grandes conjuntos de datos.
   c) La garantía de la privacidad absoluta de los datos procesados.
   d) La reducción de los costos operativos a cero.

9. En un grafo una comunidad es:
   a) Un conjunto de nodos que no comparten ninguna similitud estructural o funcional entre sí.
   b) Un conjunto de nodos aislados.
   c) Un conjunto de nodos altamente conectados que forman un subgrafo completamente independiente del resto de la red.
   d) Un conjunto de nodos que están más densamente interconectados entre sí que con los nodos fuera del conjunto.

10. La afirmación que mejor refleja el principio subyacente de PageRank, considerando su importancia en la clasificación de los sitios web, es:
    a) PageRank valora más la cantidad de enlaces entrantes a una página web, independientemente de la calidad o relevancia de estos enlaces.
    b) La efectividad de PageRank se basa exclusivamente en el análisis de las palabras clave contenidas en los enlaces entrantes, sin considerar la estructura de enlace de la Web.
    c) El algoritmo de PageRank considera tanto la cantidad como la calidad de los enlaces entrantes, asignando mayor valor a los enlaces provenientes de sitios web considerados como "importantes".
    d) PageRank opera bajo el supuesto de que los enlaces entrantes y salientes tienen el mismo impacto en la valoración de la relevancia de una página web.

11. Analizar una red permite:
    a) Detectar posibles tendencias antes de que se conviertan en tendencia.
    b) Obtener predicciones exactas de eventos futuros en mercados financieros.
    c) Evaluar la calidad del contenido de un sitio web.
    d) Encontrar nodos "sensibles" o críticos para la red.

12. El conjunto frontera de URLs en un Web Crawler:
    a) Delimita los sitios web que visitará en futuras iteraciones del proceso.
    b) Es similar a un conjunto de URLs que esperan ser visitadas.
    c) Almacena hipervínculos que pertenecen al mismo dominio del conjunto semilla de URLs con que inició el crawler.
    d) Indica las secciones que no pueden ser visitadas de cada sitio web.

13. La Web 3 se conoce como:
    a) Internet de las cosas.
    b) Web de solo lectura.
    c) Web semántica.
    d) Web de escritura-lectura.

14. Para la RI, el análisis de las redes puede:
    a) Ayudar a identificar grupos de interés.
    b) Indicar la importancia de una entidad en la transmisión de la información.
    c) Ayudar a comprender la conectividad y la accesibilidad entre las entidades.
    d) Revelar patrones de influencia dentro de una comunidad.

15. ¿Qué algoritmos permiten obtener información de una red?
    a) Índices de centralidad.
    b) Detección de comunidades.
    c) Hypertext Induced Topic Selection (HITS).
    d) PageRank.

16. Al realizar la optimización de contenido para SEO debe considerarse:
    a) Seleccionar las palabras clave al azar.
    b) Utilizar etiquetas de título y meta descripciones únicas y relevantes para cada página.
    c) Incluir palabras clave de manera excesiva en el contenido para mejorar el posicionamiento.
    d) Crear contenido valioso y original que satisfaga las necesidades de los usuarios.

17. En un sistema de control de tráfico urbano basado en reglas, ¿cuál de las siguientes reglas sería más efectiva para manejar situaciones de congestión en una intersección?
    a) Si hay pocos vehículos en la intersección, reducir el tiempo de los semáforos en verde.
    b) Si hay muchos vehículos en la intersección, aumentar el tiempo de los semáforos en verde.
    c) Si hay un vehículo de emergencia en la intersección, detener todos los demás vehículos.

d) Si hay muchos peatones cruzando la intersección, reducir el tiempo de los semáforos en rojo.

18. En el modelo de representación del conocimiento basado en herencia se puede asegurar que:
    a) La cantidad de padres que puede tener un nodo no es mayor que 4.
    b) Las conclusiones pueden ser canceladas si el grafo es ambiguo.
    c) Solo se usa en entornos referentes a la biología.
    d) El razonamiento deducido está respaldado por al menos un camino dentro del grafo.

19. En una empresa comercial, se tiene una red donde los nodos corresponden a los empleados y las aristas representan la frecuencia con la que colaboran en las ventas. Se busca mejorar la colaboración entre los empleados para aumentar las ventas totales, por lo que la directiva debe:
    a) Utilizar la centralidad de cercanía para identificar los empleados menos cercanos a otros en la red y asignarles tareas individuales para evitar posibles conflictos y desacuerdos en el proceso de colaboración.
    b) Utilizar la centralidad de cercanía para identificar los empleados más cercanos a otros en la red y promover la colaboración entre ellos, facilitando así la comunicación y el intercambio de conocimientos para mejorar las ventas.
    c) No considerar la centralidad de cercanía de los empleados al diseñar estrategias de colaboración, ya que todos los empleados tienen el mismo acceso a la información y los recursos.
    d) No tener en cuenta la centralidad de cercanía de los empleados al diseñar estrategias de colaboración, ya que esto podría introducir complicaciones adicionales en el proceso de trabajo.

20. La premisa básica del algoritmo de PageRank para clasificar páginas web en los resultados de búsqueda es:
    a) La antigüedad de la página web es el principal factor para determinar su clasificación.
    b) El contenido y la relevancia de las palabras clave en la página web determinan su posición.
    c) Los enlaces entrantes a una página web desde otras páginas contribuyen a su importancia y clasificación.
    d) La cantidad de visitas que recibe una página web determina su clasificación.

Trabajo de Control
Sistemas de Recuperación de Información
Fecha: 15 de julio de 2024
Temario #19

Nombre: _____ Grupo: _____

Seleccione en cada caso la(s) respuesta(s) correcta(s). Considere también no marcar ninguna opción.

1. Una buena práctica de SEO para mejorar el posicionamiento de un sitio web en los motores de búsqueda es:
   a) Obtener enlaces de otros sitios web relevantes y de calidad que apunten al sitio.
   b) Copiar contenido directamente de otros sitios web populares para aumentar la cantidad de páginas indexadas.
   c) Utilizar técnicas de encubrimiento para mostrar contenido diferente a lo indexado por los motores de búsqueda.
   d) Llenar el contenido con palabras clave irrelevantes para aumentar la densidad de palabras clave.

2. ¿Qué algoritmos permiten obtener información de una red?
   a) Índices de centralidad.
   b) Detección de comunidades.
   c) Hypertext Induced Topic Selection (HITS).
   d) PageRank.

3. En el contexto de la representación del conocimiento basada en herencia, ¿qué caracteriza a la herencia cancelable?
   a) La herencia es el resultado del razonamiento no transitivo.
   b) Las conclusiones no están determinadas y dependen del nodo de interés.
   c) Las propiedades heredadas siempre se mantienen y no pueden anularse.
   d) No existe ambigüedad en las conclusiones obtenidas.

4. La centralidad de intermediación de un nodo indica:
   a) La cantidad de nodos vecinos directos.
   b) La resistencia del nodo a fallos.
   c) El número total de conexiones entrantes y salientes.
   d) La frecuencia con la que un nodo actúa como puente en el camino más corto entre otros dos nodos.

5. El algoritmo de PageRank puede describirse como un procedimiento utilizado para:
   a) Determinar la velocidad de carga de una página web en un navegador.
   b) Calcular la relevancia de una página web en función de la cantidad y calidad de los enlaces que apuntan hacia ella.
   c) Prevenir el spam y el contenido no deseado en las páginas web.
   d) Clasificar las páginas web en función de su edad y autoridad.

6. En un sistema de control de tráfico urbano basado en reglas, ¿cuál de las siguientes reglas sería más efectiva para manejar situaciones de congestión en una intersección?
   a) Si hay pocos vehículos en la intersección, reducir el tiempo de los semáforos en verde.
   b) Si hay muchos vehículos en la intersección, aumentar el tiempo de los semáforos en verde.
   c) Si hay un vehículo de emergencia en la intersección, detener todos los demás vehículos.
   d) Si hay muchos peatones cruzando la intersección, reducir el tiempo de los semáforos en rojo.

7. En una red de transporte donde cada nodo es una parada de autobús y las aristas representan si existe un carro que pasa por ambos sitios, ¿qué puede mejorar el sistema de transporte?
   a) No considerar la centralidad de intermediación de las estaciones de transporte público al planificar rutas y horarios, ya que no tiene impacto en las conexiones entre las paradas.
   b) Utilizar el grafo inducido de los nodos con mayor valor en la centralidad de grado para aplicar la centralidad de intermediación con el propósito de reforzar las paradas con mayor tráfico.
   c) Utilizar la centralidad de intermediación para identificar las paradas de transferencia clave y establecer nuevas conexiones entre diferentes líneas de transporte público.
   d) No implementar sistemas de información en tiempo real para los usuarios, ya que pueden aumentar la carga de los trabajadores.

8. La Web 2.0 se caracteriza por:
   a) La existencia de sitios web dinámicos e interactivos que permiten a los usuarios participar, comentar e interactuar tanto con los creadores de contenido como con otros usuarios.

b) El uso de distintas tecnologías para crear experiencias web más interactivas y con mayor capacidad de respuesta.

c) Las actualizaciones del contenido de los sitios web se efectúa de forma manual dentro del código HTML

d) Las plataformas ofrecen experiencias a la medida, permitiendo a los usuarios personalizar sus perfiles, recibir recomendaciones ajustadas al contenido y participar en filtrado colaborativo.

9. En la RI en el contexto de Big Data, se puede asegurar que:

a) La indexación distribuida divide los datos en múltiples fragmentos que se almacenan en varios nodos para permitir búsquedas paralelas y mejorar la escalabilidad.

b) La indexación distribuida no ofrece ventajas en términos de rendimiento y escalabilidad en comparación con la indexación centralizada.

c) La indexación distribuida es una técnica obsoleta.

d) La indexación distribuida almacena todos los datos en un solo servidor para facilitar su acceso y búsqueda.

10. La diferencia entre la indexación por tokens y la indexación por conceptos puede definirse como:

a) La indexación por tokens asigna pesos a los términos basados en su importancia relativa, mientras que la indexación por conceptos utiliza un sistema de etiquetado para asociar términos con características generales.

b) La indexación por tokens divide los datos en términos individuales, mientras que la indexación por conceptos agrupa los datos en categorías definidas.

c) La indexación por tokens asigna un valor numérico a cada término de los datos, mientras que la indexación por conceptos utiliza algoritmos de encriptación para proteger la privacidad de los datos.

d) La indexación por tokens normaliza los datos reduciéndolos a su forma básica, mientras que la indexación por conceptos utiliza un método de ordenación para organizar los términos característicos de los datos.

11. Considerando las dimensiones y desafíos inherentes a Big Data puede afirmarse, tomando en cuenta las características clave y las implicaciones para su procesamiento y análisis, que:

a) Big Data se caracteriza principalmente por su pequeño volumen y uniformidad, permitiendo un procesamiento eficiente con mínimas adaptaciones de las herramientas de análisis de datos tradicionales.

b) Big Data no desafía la capacidad de las herramientas tradicionales de procesamiento de datos para capturar, almacenar, gestionar y analizar efectivamente la información, dada la evolución constante de las capacidades computacionales y algoritmos de optimización.

c) Aunque Big Data puede incluir datos estructurados, su naturaleza se expande al incorporar grandes cantidades de datos no estructurados y semiestructurados, lo que exige el uso de tecnologías especializadas en almacenamiento y procesamiento como Hadoop y sistemas de bases de datos NoSQL.

d) Además de su complejidad y diversidad, Big Data introduce desafíos significativos en términos de veracidad y variabilidad al requerir métodos avanzados de limpieza y validación de datos para asegurar la integridad del análisis.

12. El concepto de *rank sink* en el algoritmo de PageRank representa páginas web:

a) Con distribuciones uniformes de la puntuación de PageRank.

b) Que tienen una puntuación de PageRank más alta que otras debido a la manipulación de enlaces entrantes y salientes.

c) Con una baja calidad de contenido y una cantidad insuficiente de enlaces salientes, lo que las hace menos relevantes en los resultados de búsqueda.

d) Con un alto número de enlaces salientes que no reciben enlaces entrantes, lo que puede afectar negativamente su puntuación de PageRank.

13. El propósito de la política de revisitado en los Web Crawlers es:

a) Determinar la frecuencia con la que los crawlers deben volver a visitar una URL específica para mantener la información actualizada en el índice de búsqueda.

b) Establecer reglas sobre el tiempo máximo que los crawlers pueden pasar en un sitio web durante cada visita para evitar sobrecargar los servidores.

c) Limitar el acceso de los crawlers a ciertos servidores luego de visitar las páginas alojadas en estos.

d) Definir la prioridad de rastreo de las URLs, determinando el orden en que los crawlers visitan y procesan cada página web.

14. Con respecto a WordNet y su contribución a las ontologías y la representación del conocimiento en los sistemas de RI, puede afirmarse que:

a) En WordNet, una palabra está asociada a un conjunto de sinónimos (synsets), siendo estas palabras intercambiables en un contexto.

b) La integración de WordNet en sistemas de RI permite la expansión de consultas y la mejora de la precisión de los resultados al entender mejor el significado de los términos de búsqueda a través de su contexto semántico y las relaciones entre palabras.

c) WordNet diferencia claramente entre los significados de palabras según su uso en diferentes contextos, lo que permite aplicaciones avanzadas en desambiguación semántica más allá de los sistemas de recomendación y RI.

d) Una limitación de WordNet en la representación del conocimiento es su enfoque en el idioma inglés, lo que plantea desafíos en la aplicación global y la interoperabilidad con sistemas de información multilingües.

15. En un sistema donde el conocimiento está orientado a objetos se puede asegurar que:

a) Se enfatiza la atención a la información de la cual se extrajo el conocimiento.

b) Los marcos y las bandas son las estructuras utilizadas para la representación del modelo.

c) No es posible definir especializaciones de los objetos de la vida real dentro del sistema.

d) Las funciones de agregación dificultan poder establecer relaciones entre los objetos.

16. En un SRI la indexación:

a) Consiste en asociar un identificador único a cada dato almacenado en el sistema.

b) Mejora la experiencia del usuario.

c) Permite la organización y la categorización de la información.

d) Facilita la RI relevante.

17. ¿Qué es un "Uniform Resource Locator (URL) Frontier" en el contexto de Web Crawling?

a) Un protocolo que define cómo se deben formatear las URLs para el crawling.

b) Una lista prioritaria de URLs que aún no han sido visitadas por el crawler.

c) Una técnica para filtrar URLs irrelevantes y mejorar la eficiencia del crawling.

d) Una base de datos que almacena URLs únicas identificadas como recursos en la web.

18. Se tiene una red de ingredientes donde cada uno representa un nodo y las aristas simbolizan que los ingredientes forman parte de una misma receta. Se busca mejorar la experiencia culinaria mediante la elaboración de combinaciones de ingredientes más interesantes y creativas, para lo cual se debe:

a) No considerar la centralidad de grado de los ingredientes, ya que todas las combinaciones de ingredientes son igualmente válidas.

b) Utilizar la centralidad de grado para identificar los ingredientes menos conectados en la red y tomarlos en cuenta para su inclusión en las combinaciones.

c) No tener en cuenta la centralidad de grado de los ingredientes para no darle mayor importancia a los ingredientes más comunes.

d) Utilizar la centralidad de grado para identificar los ingredientes más populares en la red y crear combinaciones que incluyan una variedad de ingredientes menos comunes.

19. Dentro del análisis de redes, la centralidad de grado mide la importancia de un nodo basándose en:

a) El grado del nodo.

b) La cantidad de veces que aparece el nodo en el camino mínimo entre cualquier par de nodos.

c) La cantidad de vecinos del nodo.

d) El número de aristas que posee el nodo.

20. El uso de Hadoop y MapReduce en el contexto de la RI tiene como objetivo:

a) Automatizar el mantenimiento de sistemas de bases de datos.

b) Facilitar el análisis en tiempo real de datos de redes sociales.

c) Mejorar la eficiencia energética en centros de datos.

d) Procesar y analizar grandes conjuntos de datos para la RI.

Trabajo de Control
Sistemas de Recuperación de Información
Fecha: 15 de julio de 2024
Temario #20

Nombre: _____ Grupo: _____

Seleccione en cada caso la(s) respuesta(s) correcta(s). Considere también no marcar ninguna opción.

1. Dentro del ecosistema de Hadoop, el HDFS se caracteriza por:
   a) El modelo de acceso y de escritura de datos en tiempo real.
   b) La tolerancia a fallos mediante la replicación de datos.
   c) El almacenamiento exclusivo para archivos de texto.
   d) La capacidad ilimitada de almacenamiento.

2. Existen varias estrategias consideradas como no recomendables para mejorar el SEO de un sitio web. Dentro de estas estrategias negativas, se encuentran:
   a) Construir enlaces naturales de sitios web con autoridad y relevancia temática.
   b) Evitar el uso de las meta etiquetas y descripciones del sitio web para reflejar el contenido de la página.
   c) Incluir una cantidad excesiva de palabras clave irrelevantes para intentar manipular los rankings de búsqueda.
   d) Utilizar técnicas de encubrimiento para mostrar contenido diferente a lo indexado por los motores de búsqueda.

3. Un sistema cuenta con la siguiente información:
   - Juan pertenece a la marina.
   - Juan es capellán (sacerdote encargado de una tarea específica fuera de la parroquia).
   - Los infantes de la marina suelen ser bebedores de cerveza.
   - Un capellán no suele ser bebedor de cerveza.
   - Un bebedor de cerveza suele tener sobrepeso.
   - Por lo general, un infante de la marina está en buena condición física.

   Si se representa la información en una red de herencia se puede concluir que:
   a) No se puede asegurar que Juan sea capellán.
   b) No se puede asegurar que Juan esté en buena condición física.
   c) Juan tiene sobrepeso.
   d) Existen dos razonamientos cancelables.

4. Una empresa de comercio electrónico necesita mejorar su motor de búsqueda para proporcionar resultados más relevantes a sus usuarios. Actualmente, los resultados de la búsqueda no son precisos y los usuarios a menudo encuentran dificultades para encontrar productos específicos. La empresa está considerando implementar una indexación por conceptos para mejorar la relevancia de los resultados de búsqueda. ¿Qué beneficios podría provocar este cambio?
   a) Permite adaptarse fácilmente a cambios en el vocabulario y la terminología utilizada en las descripciones de los productos.
   b) Ayuda a identificar automáticamente relaciones entre productos, lo que puede mejorar las recomendaciones personalizadas a los usuarios.
   c) Facilita la visualización de productos al agruparlos por categorías o características comunes.
   d) Mejora la precisión en la búsqueda de productos relacionados, incluso cuando no coinciden exactamente con los términos de búsqueda del usuario.

5. En el contexto de la RI en redes, se puede afirmar que:
   a) La detección de comunidades no es relevante para la RI en redes, ya que se centra únicamente en la estructura de la red sin considerar el contenido.
   b) La detección de comunidades en una red siempre produce resultados objetivos y consistentes independientemente del algoritmo utilizado.
   c) La detección de comunidades ayuda a identificar grupos de nodos altamente conectados entre sí, lo que puede ser útil para comprender la estructura y el contenido de la red.
   d) La detección de comunidades solo se aplica a redes pequeñas y simples, no a redes grandes y complejas.

6. Los algoritmos para detectar comunidades en una red intentan:
   a) Buscar conjuntos donde cada nodo de un mismo conjunto tenga características similares al resto de los nodos del conjunto.
   b) Seleccionar aleatoriamente nodos de alto grado y sus vecinos.
   c) Buscar subgrafos tal que no incluyan nodos cuya ausencia desconecte al subgrafo.

d) Encontrar grupos donde los nodos pertenecientes a los mismos grupos son cercanos bajo cierta métrica y lejanos con respecto a los nodos de otros grupos.

7. En un sistema donde el conocimiento está definido a partir de reglas se puede asegurar que:
   a) El nivel de especificidad de las reglas está limitado.
   b) La representación de conocimiento está basada en la lógica proposicional.
   c) El razonamiento se activa en cadena hacia delante.
   d) El orden en que se definen las reglas no altera el razonamiento del sistema.

8. Sobre WordNet y su aplicación en el procesamiento del lenguaje natural, se puede afirmar que:
   a) El diseño de WordNet facilita su integración en aplicaciones multilingües de NLP, aunque su desarrollo original se centró en el inglés.
   b) Aunque WordNet es una herramienta valiosa en el NLP, su estructura no incluye información sobre la frecuencia de uso de las palabras en el lenguaje natural.
   c) Los synsets facilitan la identificación de relaciones semánticas entre palabras, como la hiperonimia y la meronimia, enriqueciendo tareas de NLP.
   d) WordNet proporciona una base para la desambiguación semántica al agrupar palabras con significados similares en synsets.

9. Si se tiene el conjunto de páginas interconectadas $\{A \to B, C; B \to C; C \to A\}$, entonces la página con valor más alto de PageRank es:
   a) A y B.
   b) A.
   c) C.
   d) B.

10. Una plataforma de comercio electrónico desea mejorar la experiencia del usuario al permitir una navegación más personalizada y contextualizada. Actualmente los usuarios tienen dificultades para encontrar productos relevantes debido a la gran cantidad de opciones disponibles. La empresa está interesada en implementar características de la Web 2.5 y la Web Semántica para abordar este problema. ¿Qué características podrían ayudar para ofrecer una navegación más personalizada y contextualizada?
    a) La implementación de ontologías y metadatos para enriquecer la descripción de productos y mejorar la precisión de las recomendaciones.
    b) La optimización de la velocidad de carga del sitio web para mejorar la experiencia del usuario y reducir el abandono del carrito de compra.
    c) La integración de redes sociales para permitir la recomendación de productos basada en las preferencias de amigos y contactos.
    d) La incorporación de los productos en tendencia en el mercado.

11. Para transformar el contenido no estructurado de las páginas web en datos estructurados, el Web Scraping:
    a) Extrae información basada en patrones de HTML/CSS.
    b) Analiza los protocolos de red.
    c) Interpreta el código JavaScript en tiempo real.
    d) Convierte de forma automática imágenes a texto.

12. ¿Cuál de las siguientes opciones describe mejor la diferencia clave entre Web Crawling y Web Scraping?
    a) Web Crawling se centra en la exploración y recopilación de enlaces de múltiples sitios web, mientras que Web Scraping se enfoca en la extracción específica de datos de páginas web individuales.
    b) Web Crawling se realiza utilizando herramientas de automatización como Selenium WebDriver, mientras que Web Scraping se lleva a cabo mediante el análisis de HTML y CSS.
    c) Web Scraping implica el análisis de la estructura y el contenido de las páginas web para extraer datos, mientras que Web Crawling se refiere a la descarga y almacenamiento de páginas web completas.
    d) Web Scraping es más eficaz para rastrear e indexar contenido web para motores de búsqueda, mientras que Web Crawling se utiliza principalmente para la extracción de datos en proyectos de investigación.

13. Considerando las dimensiones y desafíos inherentes a Big Data puede afirmarse, tomando en cuenta las características clave y las implicaciones para su procesamiento y análisis, que:
    a) Big Data se caracteriza principalmente por su pequeño volumen y uniformidad, permitiendo un procesamiento eficiente con mínimas adaptaciones de las herramientas de análisis de datos tradicionales.
    b) Big Data no desafía la capacidad de las herramientas tradicionales de procesamiento de datos para capturar, almacenar, gestionar y analizar efectivamente la información, dada la evolución constante de las capacidades computacionales y algoritmos de optimización.
    c) Aunque Big Data puede incluir datos estructurados, su naturaleza se expande al incorporar grandes cantidades de datos no estructurados y semiestructurados, lo que exige el uso de tecnologías especializadas en almacenamiento y procesamiento como Hadoop y sistemas de bases de datos NoSQL.

d) Además de su complejidad y diversidad, Big Data introduce desafíos significativos en términos de veracidad y variabilidad al requerir métodos avanzados de limpieza y validación de datos para asegurar la integridad del análisis.

14. ¿Qué es un "Uniform Resource Locator (URL) Frontier" en el contexto de Web Crawling?
    a) Un protocolo que define cómo se deben formatear las URLs para el crawling.
    b) Una lista prioritaria de URLs que aún no han sido visitadas por el crawler.
    c) Una técnica para filtrar URLs irrelevantes y mejorar la eficiencia del crawling.
    d) Una base de datos que almacena URLs únicas identificadas como recursos en la web.

15. En una red de transporte donde cada nodo es una parada de autobús y las aristas representan si existe un carro que pasa por ambos sitios, ¿qué puede mejorar el sistema de transporte?
    a) No considerar la centralidad de intermediación de las estaciones de transporte público al planificar rutas y horarios, ya que no tiene impacto en las conexiones entre las paradas.
    b) Utilizar el grafo inducido de los nodos con mayor valor en la centralidad de grado para aplicar la centralidad de intermediación con el propósito de reforzar las paradas con mayor tráfico.
    c) Utilizar la centralidad de intermediación para identificar las paradas de transferencia clave y establecer nuevas conexiones entre diferentes líneas de transporte público.
    d) No implementar sistemas de información en tiempo real para los usuarios, ya que pueden aumentar la carga de los trabajadores.

16. En una biblioteca digital se necesita implementar un sistema que permita a los usuarios encontrar libros y artículos científicos de forma rápida y precisa. Los documentos están en diversos formatos, incluyendo PDF, EPUB y HTML. Se requiere seleccionar un algoritmo de indexación adecuado para el sistema, por lo que el programador designado para la implementación debe considerar:
    a) La velocidad de indexación y recuperación de datos.
    b) La capacidad para manejar documentos en diferentes formatos.
    c) La complejidad del algoritmo en términos de implementación y mantenimiento.
    d) La capacidad del algoritmo para procesar imágenes incrustadas o referenciadas en los ficheros.

17. Si una red cumple la propiedad de ser un grafo de mundo pequeño, entonces se conoce que:
    a) El número de componentes fuertemente conexas está relacionado con la cantidad de grafos K_n presentes.
    b) La longitud media del camino entre todo par de vértices es pequeña.
    c) El grafo es un anillo regular de grado 5.
    d) La red posee pocos vértices.

18. Sobre el algoritmo de PageRank, visto en clase, se puede afirmar que:
    a) Evalúa la importancia de un sitio web en función de la calidad y cantidad de enlaces entrantes que recibe de otros sitios web.
    b) Solo tiene en cuenta el contenido en un sitio web para determinar su relevancia en los resultados de búsqueda.
    c) Asigna una puntuación alta a los sitios web que tienen un gran número de enlaces entrantes sin tener en cuenta la calidad de esos enlaces.
    d) Asigna una puntuación baja a los sitios web que contienen muchos enlaces salientes, ya que indica una falta de relevancia.

19. La premisa básica del algoritmo de PageRank para clasificar páginas web en los resultados de búsqueda es:
    a) La antigüedad de la página web es el principal factor para determinar su clasificación.
    b) El contenido y la relevancia de las palabras clave en la página web determinan su posición.
    c) Los enlaces entrantes a una página web desde otras páginas contribuyen a su importancia y clasificación.
    d) La cantidad de visitas que recibe una página web determina su clasificación.

20. El uso de Hadoop y MapReduce en el contexto de la RI tiene como objetivo:
    a) Automatizar el mantenimiento de sistemas de bases de datos.
    b) Facilitar el análisis en tiempo real de datos de redes sociales.
    c) Mejorar la eficiencia energética en centros de datos.
    d) Procesar y analizar grandes conjuntos de datos para la RI.

Trabajo de Control
Sistemas de Recuperación de Información
Fecha: 15 de julio de 2024
Temario #21

Nombre: _____ Grupo: _____

Seleccione en cada caso la(s) respuesta(s) correcta(s). Considere también no marcar ninguna opción.

1. En una empresa comercial, se tiene una red donde los nodos corresponden a los empleados y las aristas representan la frecuencia con la que colaboran en las ventas. Se busca mejorar la colaboración entre los empleados para aumentar las ventas totales, por lo que la directiva debe:
   a) Utilizar la centralidad de cercanía para identificar los empleados menos cercanos a otros en la red y asignarles tareas individuales para evitar posibles conflictos y desacuerdos en el proceso de colaboración.
   b) Utilizar la centralidad de cercanía para identificar los empleados más cercanos a otros en la red y promover la colaboración entre ellos, facilitando así la comunicación y el intercambio de conocimientos para mejorar las ventas.
   c) No considerar la centralidad de cercanía de los empleados al diseñar estrategias de colaboración, ya que todos los empleados tienen el mismo acceso a la información y los recursos.
   d) No tener en cuenta la centralidad de cercanía de los empleados al diseñar estrategias de colaboración, ya que esto podría introducir complicaciones adicionales en el proceso de trabajo.

2. Considerando las prácticas éticas y legales en el Web Scraping, se puede asegurar que:
   a) Web Scraping requiere considerar las políticas de `robots.txt` del sitio web objetivo.
   b) Es importante revisar y respetar los términos de servicio del sitio web, así como las leyes aplicables de protección de datos y derechos de autor, antes de realizar Web Scraping.
   c) Web Scraping sobre datos personales sin consentimiento es generalmente aceptado si los datos se utilizan con fines de investigación.
   d) La extracción de datos mediante Web Scraping siempre es legal, independientemente de las leyes locales sobre derechos de autor y privacidad de datos definidos en los sitios web.

3. La Web 1.0 se caracteriza por:
   a) Las actualizaciones del contenido de los sitios web se efectúa de forma manual dentro del código HTML
   b) Los sitios web se centran en brindar información en lugar de facilitar la colaboración o participación de los usuarios.
   c) Uso de comunidades virtuales para popularizar los sitios web de noticias.
   d) Los propietarios de los sitios web proporcionan contenido de forma periódica.

4. En una red de transporte donde cada nodo es una parada de autobús y las aristas representan si existe un carro que pasa por ambos sitios, ¿qué puede mejorar el sistema de transporte?
   a) No considerar la centralidad de intermediación de las estaciones de transporte público al planificar rutas y horarios, ya que no tiene impacto en las conexiones entre las paradas.
   b) Utilizar el grafo inducido de los nodos con mayor valor en la centralidad de grado para aplicar la centralidad de intermediación con el propósito de reforzar las paradas con mayor tráfico.
   c) Utilizar la centralidad de intermediación para identificar las paradas de transferencia clave y establecer nuevas conexiones entre diferentes líneas de transporte público.
   d) No implementar sistemas de información en tiempo real para los usuarios, ya que pueden aumentar la carga de los trabajadores.

5. Al realizar la optimización de contenido para SEO debe considerarse:
   a) Seleccionar las palabras clave al azar.
   b) Utilizar etiquetas de título y meta descripciones únicas y relevantes para cada página.
   c) Incluir palabras clave de manera excesiva en el contenido para mejorar el posicionamiento.
   d) Crear contenido valioso y original que satisfaga las necesidades de los usuarios.

6. La diferencia entre la indexación por tokens y la indexación por conceptos puede definirse como:
   a) La indexación por tokens asigna pesos a los términos basados en su importancia relativa, mientras que la indexación por conceptos utiliza un sistema de etiquetado para asociar términos con características generales.

b) La indexación por tokens divide los datos en términos individuales, mientras que la indexación por conceptos agrupa los datos en categorías definidas.

c) La indexación por tokens asigna un valor numérico a cada término de los datos, mientras que la indexación por conceptos utiliza algoritmos de encriptación para proteger la privacidad de los datos.

d) La indexación por tokens normaliza los datos reduciéndolos a su forma básica, mientras que la indexación por conceptos utiliza un método de ordenación para organizar los términos característicos de los datos.

7. Una empresa de comercio electrónico necesita mejorar su motor de búsqueda para proporcionar resultados más relevantes a sus usuarios. Actualmente, los resultados de la búsqueda no son precisos y los usuarios a menudo encuentran dificultades para encontrar productos específicos. La empresa está considerando implementar una indexación por conceptos para mejorar la relevancia de los resultados de búsqueda. ¿Qué beneficios podría provocar este cambio?

a) Permite adaptarse fácilmente a cambios en el vocabulario y la terminología utilizada en las descripciones de los productos.

b) Ayuda a identificar automáticamente relaciones entre productos, lo que puede mejorar las recomendaciones personalizadas a los usuarios.

c) Facilita la visualización de productos al agruparlos por categorías o características comunes.

d) Mejora la precisión en la búsqueda de productos relacionados, incluso cuando no coinciden exactamente con los términos de búsqueda del usuario.

8. ¿Qué estrategia utilizan los Web Crawlers para asegurar un rastreo eficiente y respetuoso de los recursos de los sitios web?

a) Ignorar completamente el archivo `robots.txt` de los sitios web.

b) Visitar y rastrear todos los enlaces de una página web simultáneamente.

c) Extraer únicamente contenido multimedia para reducir la carga en los servidores web.

d) Seguir las directrices del archivo `robots.txt` y aplicar un retraso entre las solicitudes.

9. En una biblioteca digital se necesita implementar un sistema que permita a los usuarios encontrar libros y artículos científicos de forma rápida y precisa. Los documentos están en diversos formatos, incluyendo PDF, EPUB y HTML. Se requiere seleccionar un algoritmo de indexación adecuado para el sistema, por lo que el programador designado para la implementación debe considerar:

a) La velocidad de indexación y recuperación de datos.

b) La capacidad para manejar documentos en diferentes formatos.

c) La complejidad del algoritmo en términos de implementación y mantenimiento.

d) La capacidad del algoritmo para procesar imágenes incrustadas o referenciadas en los ficheros.

10. En el modelo de representación del conocimiento basado en herencia se puede asegurar que:

a) La cantidad de padres que puede tener un nodo no es mayor que 4.

b) Las conclusiones pueden ser canceladas si el grafo es ambiguo.

c) Solo se usa en entornos referentes a la biología.

d) El razonamiento deducido está respaldado por al menos un camino dentro del grafo.

11. Considerando las dimensiones y desafíos inherentes a Big Data puede afirmarse, tomando en cuenta las características clave y las implicaciones para su procesamiento y análisis, que:

a) Big Data se caracteriza principalmente por su pequeño volumen y uniformidad, permitiendo un procesamiento eficiente con mínimas adaptaciones de las herramientas de análisis de datos tradicionales.

b) Big Data no desafía la capacidad de las herramientas tradicionales de procesamiento de datos para capturar, almacenar, gestionar y analizar efectivamente la información, dada la evolución constante de las capacidades computacionales y algoritmos de optimización.

c) Aunque Big Data puede incluir datos estructurados, su naturaleza se expande al incorporar grandes cantidades de datos no estructurados y semiestructurados, lo que exige el uso de tecnologías especializadas en almacenamiento y procesamiento como Hadoop y sistemas de bases de datos NoSQL.

d) Además de su complejidad y diversidad, Big Data introduce desafíos significativos en términos de veracidad y variabilidad al requerir métodos avanzados de limpieza y validación de datos para asegurar la integridad del análisis.

12. En la RI en el contexto de Big Data, se puede asegurar que:

a) La indexación distribuida divide los datos en múltiples fragmentos que se almacenan en varios nodos para permitir búsquedas paralelas y mejorar la escalabilidad.

b) La indexación distribuida no ofrece ventajas en términos de rendimiento y escalabilidad en comparación con la indexación centralizada.

c) La indexación distribuida es una técnica obsoleta.

d) La indexación distribuida almacena todos los datos en un solo servidor para facilitar su acceso y búsqueda.

13. Al "relajar" el concepto de clique en la detección de comunidades se intenta solucionar:
    a) La imposibilidad de conectar nodos distantes.
    b) La necesidad de datos externos para analizar la red.
    c) La uniformidad de los nodos en términos de grado.
    d) El solapamiento y la complejidad computacional.

14. En el contexto de la representación del conocimiento basada en herencia, ¿qué caracteriza a la herencia cancelable?
    a) La herencia es el resultado del razonamiento no transitivo.
    b) Las conclusiones no están determinadas y dependen del nodo de interés.
    c) Las propiedades heredadas siempre se mantienen y no pueden anularse.
    d) No existe ambigüedad en las conclusiones obtenidas.

15. El algoritmo de PageRank puede describirse como un procedimiento utilizado para:
    a) Determinar la velocidad de carga de una página web en un navegador.
    b) Calcular la relevancia de una página web en función de la cantidad y calidad de los enlaces que apuntan hacia ella.
    c) Prevenir el spam y el contenido no deseado en las páginas web.
    d) Clasificar las páginas web en función de su edad y autoridad.

16. En el contexto de la RI en redes sociales se puede afirmar que:
    a) El análisis de redes sociales se centra exclusivamente en la cantidad de seguidores que tiene un usuario en particular para determinar su influencia en la red.
    b) El análisis de centralidad de intermediación se utiliza para identificar usuarios que son importantes en una red social debido a su posición como "puentes" entre diferentes grupos de usuarios.
    c) El análisis de sentimientos se utiliza para determinar la popularidad de una publicación en redes sociales sin tener en cuenta la opinión de los usuarios.
    d) El análisis de redes sociales no es útil para comprender la difusión de información en una red social específica, para ello se utiliza la medida de centralidad de vector propio.

17. En el contexto de la RI en redes, se puede afirmar que:
    a) La detección de comunidades no es relevante para la RI en redes, ya que se centra únicamente en la estructura de la red sin considerar el contenido.
    b) La detección de comunidades en una red siempre produce resultados objetivos y consistentes independientemente del algoritmo utilizado.
    c) La detección de comunidades ayuda a identificar grupos de nodos altamente conectados entre sí, lo que puede ser útil para comprender la estructura y el contenido de la red.
    d) La detección de comunidades solo se aplica a redes pequeñas y simples, no a redes grandes y complejas.

18. Dentro del ecosistema de Hadoop, el HDFS se caracteriza por:
    a) El modelo de acceso y de escritura de datos en tiempo real.
    b) La tolerancia a fallos mediante la replicación de datos.
    c) El almacenamiento exclusivo para archivos de texto.
    d) La capacidad ilimitada de almacenamiento.

19. Para la RI, el análisis de las redes puede:
    a) Ayudar a identificar grupos de interés.
    b) Indicar la importancia de una entidad en la transmisión de la información.
    c) Ayudar a comprender la conectividad y la accesibilidad entre las entidades.
    d) Revelar patrones de influencia dentro de una comunidad.

20. La web actual se enfrenta a problemas como:
    a) Presencia de grandes volúmenes de datos estructurados.
    b) Presencia de una alta calidad en los datos.
    c) Presencia de datos volátiles y distribuidos.
    d) Heterogeneidad en los datos.

Trabajo de Control
Sistemas de Recuperación de Información
Fecha: 15 de julio de 2024
Temario #22

Nombre: _____ Grupo: _____

Seleccione en cada caso la(s) respuesta(s) correcta(s). Considere también no marcar ninguna opción.

1. El uso de Hadoop y MapReduce en el contexto de la RI tiene como objetivo:
   a) Automatizar el mantenimiento de sistemas de bases de datos.
   b) Facilitar el análisis en tiempo real de datos de redes sociales.
   c) Mejorar la eficiencia energética en centros de datos.
   d) Procesar y analizar grandes conjuntos de datos para la RI.

2. En la RI en el contexto de Big Data, se puede asegurar que:
   a) El almacenamiento distribuido centraliza todos los datos en un único servidor para facilitar su gestión y mantenimiento.
   b) El almacenamiento distribuido ofrece ventajas significativas en términos de escalabilidad y rendimiento en comparación con el almacenamiento centralizado.
   c) El almacenamiento distribuido reparte los datos en múltiples servidores para mejorar la disponibilidad y la redundancia del sistema.
   d) El almacenamiento distribuido es una técnica obsoleta.

3. El posicionamiento de un sitio web en los motores de búsqueda puede ser afectado por:
   a) Utilizar técnicas de encubrimiento para mostrar contenido diferente a lo indexado por los motores de búsqueda.
   b) Contenido sin valor en el sitio web.
   c) Mantener una estructura de URL clara y coherente.
   d) Obtener enlaces de sitios web irrelevantes y de baja calidad.

4. Un SRI es capaz de:
   a) Crear los índices asociados a los datos sin tener que analizar cada dato.
   b) Generar índices invertidos de manera óptima sin considerar el contexto.
   c) Reducir el tiempo de indexación de los datos si utiliza servidores distribuidos dentro de la red para que cada uno ejecute la indexación del mismo conjunto de datos.
   d) No necesitar de ningún almacenamiento externo para alojar los índices de los datos.

5. La política de ordenación de URLs en los Web Crawlers tiene como aspecto fundamental:
   a) Limitar el acceso de los crawlers a ciertas secciones de un sitio web, evitando el rastreo de URLs consideradas menos importantes o sensibles.
   b) Definir la estructura de la URL de destino, asegurando que estén ordenadas alfabéticamente para facilitar la navegación y la indexación.
   c) Establecer la prioridad de rastreo de las URLs, determinando el orden en que los crawlers visitan y procesan cada página web.
   d) Determinar la forma en que los crawlers asignan un valor de relevancia a cada URL para clasificarlas en el índice de búsqueda.

6. Dentro del ecosistema de Hadoop, el HDFS se caracteriza por:
   a) El modelo de acceso y de escritura de datos en tiempo real.
   b) La tolerancia a fallos mediante la replicación de datos.
   c) El almacenamiento exclusivo para archivos de texto.
   d) La capacidad ilimitada de almacenamiento.

7. El algoritmo de PageRank converge si:
   a) La norma de la diferencia entre los vectores es menor a un umbral predefinido.
   b) Finaliza la ejecución cuando el número de iteraciones excede un máximo de iteraciones previamente definido.
   c) Se define un factor de normalización en la fórmula de la función.
   d) El algoritmo no se implementa de forma iterativa.

8. El algoritmo de PageRank puede describirse como un procedimiento utilizado para:
   a) Determinar la velocidad de carga de una página web en un navegador.
   b) Calcular la relevancia de una página web en función de la cantidad y calidad de los enlaces que apuntan hacia ella.
   c) Prevenir el spam y el contenido no deseado en las páginas web.

d) Clasificar las páginas web en función de su edad y autoridad.

9. Con respecto a WordNet y su contribución a las ontologías y la representación del conocimiento en los sistemas de RI, puede afirmarse que:
   a) En WordNet, una palabra está asociada a un conjunto de sinónimos (synsets), siendo estas palabras intercambiables en un contexto.
   b) La integración de WordNet en sistemas de RI permite la expansión de consultas y la mejora de la precisión de los resultados al entender mejor el significado de los términos de búsqueda a través de su contexto semántico y las relaciones entre palabras.
   c) WordNet diferencia claramente entre los significados de palabras según su uso en diferentes contextos, lo que permite aplicaciones avanzadas en desambiguación semántica más allá de los sistemas de recomendación y RI.
   d) Una limitación de WordNet en la representación del conocimiento es su enfoque en el idioma inglés, lo que plantea desafíos en la aplicación global y la interoperabilidad con sistemas de información multilingües.

10. Para transformar el contenido no estructurado de las páginas web en datos estructurados, el Web Scraping:
    a) Extrae información basada en patrones de HTML/CSS.
    b) Analiza los protocolos de red.
    c) Interpreta el código JavaScript en tiempo real.
    d) Convierte de forma automática imágenes a texto.

11. Una plataforma de comercio electrónico desea mejorar la experiencia del usuario al permitir una navegación más personalizada y contextualizada. Actualmente los usuarios tienen dificultades para encontrar productos relevantes debido a la gran cantidad de opciones disponibles. La empresa está interesada en implementar características de la Web 2.5 y la Web Semántica para abordar este problema. ¿Qué características podrían ayudar para ofrecer una navegación más personalizada y contextualizada?
    a) La implementación de ontologías y metadatos para enriquecer la descripción de productos y mejorar la precisión de las recomendaciones.
    b) La optimización de la velocidad de carga del sitio web para mejorar la experiencia del usuario y reducir el abandono del carrito de compra.
    c) La integración de redes sociales para permitir la recomendación de productos basada en las preferencias de amigos y contactos.
    d) La incorporación de los productos en tendencia en el mercado.

12. Existen varias estrategias consideradas como no recomendables para mejorar el SEO de un sitio web. Dentro de estas estrategias negativas, se encuentran:
    a) Construir enlaces naturales de sitios web con autoridad y relevancia temática.
    b) Evitar el uso de las meta etiquetas y descripciones del sitio web para reflejar el contenido de la página.
    c) Incluir una cantidad excesiva de palabras clave irrelevantes para intentar manipular los rankings de búsqueda.
    d) Utilizar técnicas de encubrimiento para mostrar contenido diferente a lo indexado por los motores de búsqueda.

13. Para contribuir positivamente al posicionamiento orgánico de un sitio web en los motores de búsqueda se puede:
    a) Crear contenido relevante y de alta calidad que satisfaga las necesidades de información de los usuarios.
    b) Mejorar la velocidad de carga del sitio web y asegurar que sea *responsive* y fácil de usar en dispositivos móviles.
    c) Incluir una densidad alta de palabras clave para asegurar que el sitio web aparezca en tantas búsquedas como sea posible.
    d) Obtener enlaces entrantes de otros sitios web de alta autoridad y relevancia temática.

14. En el algoritmo de Indexación Basada en Clasificación Bloqueada (BSBI), ¿cuál es el paso final para crear un índice invertido para la colección completa de los datos?
    a) Eliminar los términos duplicados de los índices.
    b) Indexar cada bloque de forma independiente.
    c) Fusionar los índices invertidos de cada bloque.
    d) Dividir la colección de datos en bloques de tamaño fijo.

15. En una red de transporte donde cada nodo es una parada de autobús y las aristas representan si existe un carro que pasa por ambos sitios, ¿qué puede mejorar el sistema de transporte?
    a) No considerar la centralidad de intermediación de las estaciones de transporte público al planificar rutas y horarios, ya que no tiene impacto en las conexiones entre las paradas.
    b) Utilizar el grafo inducido de los nodos con mayor valor en la centralidad de grado para aplicar la centralidad de intermediación con el propósito de reforzar las paradas con mayor tráfico.
    c) Utilizar la centralidad de intermediación para identificar las paradas de transferencia clave y establecer nuevas conexiones entre diferentes líneas de transporte público.

d) No implementar sistemas de información en tiempo real para los usuarios, ya que pueden aumentar la carga de los trabajadores.

16. El conjunto frontera de URLs en un Web Crawler:
    a) Delimita los sitios web que visitará en futuras iteraciones del proceso.
    b) Es similar a un conjunto de URLs que esperan ser visitadas.
    c) Almacena hipervínculos que pertenecen al mismo dominio del conjunto semilla de URLs con que inició el crawler.
    d) Indica las secciones que no pueden ser visitadas de cada sitio web.

17. En un grafo una comunidad es:
    a) Un conjunto de nodos que no comparten ninguna similitud estructural o funcional entre sí.
    b) Un conjunto de nodos aislados.
    c) Un conjunto de nodos altamente conectados que forman un subgrafo completamente independiente del resto de la red.
    d) Un conjunto de nodos que están más densamente interconectados entre sí que con los nodos fuera del conjunto.

18. El propósito de la política de revisitado en los Web Crawlers es:
    a) Determinar la frecuencia con la que los crawlers deben volver a visitar una URL específica para mantener la información actualizada en el índice de búsqueda.
    b) Establecer reglas sobre el tiempo máximo que los crawlers pueden pasar en un sitio web durante cada visita para evitar sobrecargar los servidores.
    c) Limitar el acceso de los crawlers a ciertos servidores luego de visitar las páginas alojadas en estos.
    d) Definir la prioridad de rastreo de las URLs, determinando el orden en que los crawlers visitan y procesan cada página web.

19. La Web 2.0 se caracteriza por:
    a) La existencia de sitios web dinámicos e interactivos que permiten a los usuarios participar, comentar e interactuar tanto con los creadores de contenido como con otros usuarios.
    b) El uso de distintas tecnologías para crear experiencias web más interactivas y con mayor capacidad de respuesta.
    c) Las actualizaciones del contenido de los sitios web se efectúa de forma manual dentro del código HTML
    d) Las plataformas ofrecen experiencias a la medida, permitiendo a los usuarios personalizar sus perfiles, recibir recomendaciones ajustadas al contenido y participar en filtrado colaborativo.

20. En una empresa comercial, se tiene una red donde los nodos corresponden a los empleados y las aristas representan la frecuencia con la que colaboran en las ventas. Se busca mejorar la colaboración entre los empleados para aumentar las ventas totales, por lo que la directiva debe:
    a) Utilizar la centralidad de cercanía para identificar los empleados menos cercanos a otros en la red y asignarles tareas individuales para evitar posibles conflictos y desacuerdos en el proceso de colaboración.
    b) Utilizar la centralidad de cercanía para identificar los empleados más cercanos a otros en la red y promover la colaboración entre ellos, facilitando así la comunicación y el intercambio de conocimientos para mejorar las ventas.
    c) No considerar la centralidad de cercanía de los empleados al diseñar estrategias de colaboración, ya que todos los empleados tienen el mismo acceso a la información y los recursos.
    d) No tener en cuenta la centralidad de cercanía de los empleados al diseñar estrategias de colaboración, ya que esto podría introducir complicaciones adicionales en el proceso de trabajo.

Trabajo de Control
Sistemas de Recuperación de Información
Fecha: 15 de julio de 2024
Temario #23

Nombre: _____ Grupo: _____

Seleccione en cada caso la(s) respuesta(s) correcta(s). Considere también no marcar ninguna opción.

1. Analizar una red permite:
   a) Detectar posibles tendencias antes de que se conviertan en tendencia.
   b) Obtener predicciones exactas de eventos futuros en mercados financieros.
   c) Evaluar la calidad del contenido de un sitio web.
   d) Encontrar nodos "sensibles" o críticos para la red.

2. Referente al Web Crawler puede afirmarse que:
   a) Los hipervínculos encontrados en cada sitio web que no pertenecen al dominio donde fueron encontrados se desechan, puesto que no expande el conjunto de URLs sin visitar.
   b) Las páginas visitadas no se procesan nunca más.
   c) No necesita de un conjunto inicial de URLs para recorrer la Web.
   d) No tiene como objetivo indexar y recopilar información de diferentes sitios web.

3. En una biblioteca digital se necesita implementar un sistema que permita a los usuarios encontrar libros y artículos científicos de forma rápida y precisa. Los documentos están en diversos formatos, incluyendo PDF, EPUB y HTML. Se requiere seleccionar un algoritmo de indexación adecuado para el sistema, por lo que el programador designado para la implementación debe considerar:
   a) La velocidad de indexación y recuperación de datos.
   b) La capacidad para manejar documentos en diferentes formatos.
   c) La complejidad del algoritmo en términos de implementación y mantenimiento.
   d) La capacidad del algoritmo para procesar imágenes incrustadas o referenciadas en los ficheros.

4. En una empresa comercial, se tiene una red donde los nodos corresponden a los empleados y las aristas representan la frecuencia con la que colaboran en las ventas. Se busca mejorar la colaboración entre los empleados para aumentar las ventas totales, por lo que la directiva debe:
   a) Utilizar la centralidad de cercanía para identificar los empleados menos cercanos a otros en la red y asignarles tareas individuales para evitar posibles conflictos y desacuerdos en el proceso de colaboración.
   b) Utilizar la centralidad de cercanía para identificar los empleados más cercanos a otros en la red y promover la colaboración entre ellos, facilitando así la comunicación y el intercambio de conocimientos para mejorar las ventas.
   c) No considerar la centralidad de cercanía de los empleados al diseñar estrategias de colaboración, ya que todos los empleados tienen el mismo acceso a la información y los recursos.
   d) No tener en cuenta la centralidad de cercanía de los empleados al diseñar estrategias de colaboración, ya que esto podría introducir complicaciones adicionales en el proceso de trabajo.

5. Para contribuir positivamente al posicionamiento orgánico de un sitio web en los motores de búsqueda se puede:
   a) Crear contenido relevante y de alta calidad que satisfaga las necesidades de información de los usuarios.
   b) Mejorar la velocidad de carga del sitio web y asegurar que sea *responsive* y fácil de usar en dispositivos móviles.
   c) Incluir una densidad alta de palabras clave para asegurar que el sitio web aparezca en tantas búsquedas como sea posible.
   d) Obtener enlaces entrantes de otros sitios web de alta autoridad y relevancia temática.

6. El algoritmo de PageRank puede describirse como un procedimiento utilizado para:
   a) Determinar la velocidad de carga de una página web en un navegador.
   b) Calcular la relevancia de una página web en función de la cantidad y calidad de los enlaces que apuntan hacia ella.
   c) Prevenir el spam y el contenido no deseado en las páginas web.
   d) Clasificar las páginas web en función de su edad y autoridad.

7. Un investigador necesita recopilar datos de múltiples sitios web para un estudio académico, pero se enfrenta a varios desafíos al realizar el proceso de Web Scraping de manera ética y legal. ¿Cuál de las siguientes opciones describe mejor uno de los desafíos asociados al proceso de Web Scraping?

a) La necesidad de comprender la estructura del sitio web y su código HTML para extraer los datos correctamente.

b) La disponibilidad limitada de datos en línea que se pueden extraer utilizando técnicas de Web Scraping.

c) La dificultad para encontrar herramientas de Web Scraping gratuitas y fiables.

d) La necesidad de estar montado sobre un scrawler que cumpla con todas las políticas.

8. Para la RI, el análisis de las redes puede:

a) Ayudar a identificar grupos de interés.

b) Indicar la importancia de una entidad en la transmisión de la información.

c) Ayudar a comprender la conectividad y la accesibilidad entre las entidades.

d) Revelar patrones de influencia dentro de una comunidad.

9. La web actual se enfrenta a problemas como:

a) Presencia de grandes volúmenes de datos estructurados.

b) Presencia de una alta calidad en los datos.

c) Presencia de datos volátiles y distribuidos.

d) Heterogeneidad en los datos.

10. La integración de Hadoop y MapReduce en la RI trae como ventaja:

a) La eliminación de la necesidad de sistemas de bases de datos.

b) La posibilidad de la extracción de información relevante y la generación de resultados significativos de grandes conjuntos de datos.

c) La garantía de la privacidad absoluta de los datos procesados.

d) La reducción de los costos operativos a cero.

11. Sobre el algoritmo de PageRank, visto en clase, se puede afirmar que:

a) Evalúa la importancia de un sitio web en función de la calidad y cantidad de enlaces entrantes que recibe de otros sitios web.

b) Solo tiene en cuenta el contenido en un sitio web para determinar su relevancia en los resultados de búsqueda.

c) Asigna una puntuación alta a los sitios web que tienen un gran número de enlaces entrantes sin tener en cuenta la calidad de esos enlaces.

d) Asigna una puntuación baja a los sitios web que contienen muchos enlaces salientes, ya que indica una falta de relevancia.

12. Considerando las prácticas éticas y legales en el Web Scraping, se puede asegurar que:

a) Web Scraping requiere considerar las políticas de `robots.txt` del sitio web objetivo.

b) Es importante revisar y respetar los términos de servicio del sitio web, así como las leyes aplicables de protección de datos y derechos de autor, antes de realizar Web Scraping.

c) Web Scraping sobre datos personales sin consentimiento es generalmente aceptado si los datos se utilizan con fines de investigación.

d) La extracción de datos mediante Web Scraping siempre es legal, independientemente de las leyes locales sobre derechos de autor y privacidad de datos definidos en los sitios web.

13. Considerando las dimensiones y desafíos inherentes a Big Data puede afirmarse, tomando en cuenta las características clave y las implicaciones para su procesamiento y análisis, que:

a) Big Data se caracteriza principalmente por su pequeño volumen y uniformidad, permitiendo un procesamiento eficiente con mínimas adaptaciones de las herramientas de análisis de datos tradicionales.

b) Big Data no desafía la capacidad de las herramientas tradicionales de procesamiento de datos para capturar, almacenar, gestionar y analizar efectivamente la información, dada la evolución constante de las capacidades computacionales y algoritmos de optimización.

c) Aunque Big Data puede incluir datos estructurados, su naturaleza se expande al incorporar grandes cantidades de datos no estructurados y semiestructurados, lo que exige el uso de tecnologías especializadas en almacenamiento y procesamiento como Hadoop y sistemas de bases de datos NoSQL.

d) Además de su complejidad y diversidad, Big Data introduce desafíos significativos en términos de veracidad y variabilidad al requerir métodos avanzados de limpieza y validación de datos para asegurar la integridad del análisis.

14. ¿Cuál de las siguientes opciones describe mejor la diferencia clave entre Web Crawling y Web Scraping?

a) Web Crawling se centra en la exploración y recopilación de enlaces de múltiples sitios web, mientras que Web Scraping se enfoca en la extracción específica de datos de páginas web individuales.

b) Web Crawling se realiza utilizando herramientas de automatización como Selenium WebDriver, mientras que Web Scraping se lleva a cabo mediante el análisis de HTML y CSS.

c) Web Scraping implica el análisis de la estructura y el contenido de las páginas web para extraer datos, mientras que Web Crawling se refiere a la descarga y almacenamiento de páginas web completas.

d) Web Scraping es más eficaz para rastrear e indexar contenido web para motores de búsqueda, mientras que Web Crawling se utiliza principalmente para la extracción de datos en proyectos de investigación.

15. El propósito de la política de revisitado en los Web Crawlers es:

a) Determinar la frecuencia con la que los crawlers deben volver a visitar una URL específica para mantener la información actualizada en el índice de búsqueda.

b) Establecer reglas sobre el tiempo máximo que los crawlers pueden pasar en un sitio web durante cada visita para evitar sobrecargar los servidores.

c) Limitar el acceso de los crawlers a ciertos servidores luego de visitar las páginas alojadas en estos.

d) Definir la prioridad de rastreo de las URLs, determinando el orden en que los crawlers visitan y procesan cada página web.

16. En el algoritmo de Indexación Basada en Clasificación Bloqueada (BSBI), ¿cuál es el paso final para crear un índice invertido para la colección completa de los datos?

a) Eliminar los términos duplicados de los índices.

b) Indexar cada bloque de forma independiente.

c) Fusionar los índices invertidos de cada bloque.

d) Dividir la colección de datos en bloques de tamaño fijo.

17. Un sistema cuenta con la siguiente información:

- Juan pertenece a la marina.
- Juan es capellán (sacerdote encargado de una tarea específica fuera de la parroquia).
- Los infantes de la marina suelen ser bebedores de cerveza.
- Un capellán no suele ser bebedor de cerveza.
- Un bebedor de cerveza suele tener sobrepeso.
- Por lo general, un infante de la marina está en buena condición física.

Si se representa la información en una red de herencia se puede concluir que:

a) No se puede asegurar que Juan sea capellán.

b) No se puede asegurar que Juan esté en buena condición física.

c) Juan tiene sobrepeso.

d) Existen dos razonamientos cancelables.

18. Dentro del análisis de redes, la centralidad de grado mide la importancia de un nodo basándose en:

a) El grado del nodo.

b) La cantidad de veces que aparece el nodo en el camino mínimo entre cualquier par de nodos.

c) La cantidad de vecinos del nodo.

d) El número de aristas que posee el nodo.

19. La Web 1.0 se caracteriza por:

a) Las actualizaciones del contenido de los sitios web se efectúa de forma manual dentro del código HTML

b) Los sitios web se centran en brindar información en lugar de facilitar la colaboración o participación de los usuarios.

c) Uso de comunidades virtuales para popularizar los sitios web de noticias.

d) Los propietarios de los sitios web proporcionan contenido de forma periódica.

20. El concepto de *rank sink* en el algoritmo de PageRank representa páginas web:

a) Con distribuciones uniformes de la puntuación de PageRank.

b) Que tienen una puntuación de PageRank más alta que otras debido a la manipulación de enlaces entrantes y salientes.

c) Con una baja calidad de contenido y una cantidad insuficiente de enlaces salientes, lo que las hace menos relevantes en los resultados de búsqueda.

d) Con un alto número de enlaces salientes que no reciben enlaces entrantes, lo que puede afectar negativamente su puntuación de PageRank.

Trabajo de Control
Sistemas de Recuperación de Información
Fecha: 15 de julio de 2024
Temario #24

Nombre: _____ Grupo: _____

Seleccione en cada caso la(s) respuesta(s) correcta(s). Considere también no marcar ninguna opción.

1. Si se tiene el conjunto de páginas interconectadas $\{A \to B, C; B \to C; C \to A\}$, entonces la página con valor más alto de PageRank es:
   a) A y B.
   b) A.
   c) C.
   d) B.

2. Existen varias estrategias consideradas como no recomendables para mejorar el SEO de un sitio web. Dentro de estas estrategias negativas, se encuentran:
   a) Construir enlaces naturales de sitios web con autoridad y relevancia temática.
   b) Evitar el uso de las meta etiquetas y descripciones del sitio web para reflejar el contenido de la página.
   c) Incluir una cantidad excesiva de palabras clave irrelevantes para intentar manipular los rankings de búsqueda.
   d) Utilizar técnicas de encubrimiento para mostrar contenido diferente a lo indexado por los motores de búsqueda.

3. En un sistema donde el conocimiento está orientado a objetos se puede asegurar que:
   a) Se enfatiza la atención a la información de la cual se extrajo el conocimiento.
   b) Los marcos y las bandas son las estructuras utilizadas para la representación del modelo.
   c) No es posible definir especializaciones de los objetos de la vida real dentro del sistema.
   d) Las funciones de agregación dificultan poder establecer relaciones entre los objetos.

4. Dentro del análisis de redes, la centralidad de grado mide la importancia de un nodo basándose en:
   a) El grado del nodo.
   b) La cantidad de veces que aparece el nodo en el camino mínimo entre cualquier par de nodos.
   c) La cantidad de vecinos del nodo.
   d) El número de aristas que posee el nodo.

5. Si una red cumple la propiedad de ser un grafo de mundo pequeño, entonces se conoce que:
   a) El número de componentes fuertemente conexas está relacionado con la cantidad de grafos K_n presentes.
   b) La longitud media del camino entre todo par de vértices es pequeña.
   c) El grafo es un anillo regular de grado 5.
   d) La red posee pocos vértices.

6. La Web 3 se conoce como:
   a) Internet de las cosas.
   b) Web de solo lectura.
   c) Web semántica.
   d) Web de escritura-lectura.

7. La Web 2.0 se caracteriza por:
   a) La existencia de sitios web dinámicos e interactivos que permiten a los usuarios participar, comentar e interactuar tanto con los creadores de contenido como con otros usuarios.
   b) El uso de distintas tecnologías para crear experiencias web más interactivas y con mayor capacidad de respuesta.
   c) Las actualizaciones del contenido de los sitios web se efectúa de forma manual dentro del código HTML
   d) Las plataformas ofrecen experiencias a la medida, permitiendo a los usuarios personalizar sus perfiles, recibir recomendaciones ajustadas al contenido y participar en filtrado colaborativo.

8. El algoritmo Hypertext Induced Topic Selection (HITS) intenta buscar nodos especiales. Estos son conocidos como:
   a) Autoridades.
   b) Hubs.
   c) Centrales.
   d) Sensibles.

9. Un SRI es capaz de:
   a) Crear los índices asociados a los datos sin tener que analizar cada dato.
   b) Generar índices invertidos de manera óptima sin considerar el contexto.
   c) Reducir el tiempo de indexación de los datos si utiliza servidores distribuidos dentro de la red para que cada uno ejecute la indexación del mismo conjunto de datos.

d) No necesitar de ningún almacenamiento externo para alojar los índices de los datos.

10. La transición de la Web 1.0 a la Web 2.0 se caracterizó principalmente por:
    a) El aumento en la velocidad de conexión a internet, que permitió una mejor calidad de las páginas web.
    b) La disminución de la importancia de los motores de búsqueda en la navegación web.
    c) El cambio de páginas web estáticas a dinámicas, permitiendo la interacción del usuario y la generación de contenido.
    d) La reducción en el uso de HTML y CSS en el desarrollo de sitios web.

11. El componente responsable de la gestión de recursos y planificación de tareas en Hadoop es:
    a) MapReduce.
    b) Hadoop Common.
    c) HDFS.
    d) YARN.

12. El algoritmo de PageRank converge si:
    a) La norma de la diferencia entre los vectores es menor a un umbral predefinido.
    b) Finaliza la ejecución cuando el número de iteraciones excede un máximo de iteraciones previamente definido.
    c) Se define un factor de normalización en la fórmula de la función.
    d) El algoritmo no se implementa de forma iterativa.

13. En un grafo una comunidad es:
    a) Un conjunto de nodos que no comparten ninguna similitud estructural o funcional entre sí.
    b) Un conjunto de nodos aislados.
    c) Un conjunto de nodos altamente conectados que forman un subgrafo completamente independiente del resto de la red.
    d) Un conjunto de nodos que están más densamente interconectados entre sí que con los nodos fuera del conjunto.

14. El propósito de la política de revisitado en los Web Crawlers es:
    a) Determinar la frecuencia con la que los crawlers deben volver a visitar una URL específica para mantener la información actualizada en el índice de búsqueda.
    b) Establecer reglas sobre el tiempo máximo que los crawlers pueden pasar en un sitio web durante cada visita para evitar sobrecargar los servidores.
    c) Limitar el acceso de los crawlers a ciertos servidores luego de visitar las páginas alojadas en estos.
    d) Definir la prioridad de rastreo de las URLs, determinando el orden en que los crawlers visitan y procesan cada página web.

15. El algoritmo de PageRank puede describirse como un procedimiento utilizado para:
    a) Determinar la velocidad de carga de una página web en un navegador.
    b) Calcular la relevancia de una página web en función de la cantidad y calidad de los enlaces que apuntan hacia ella.
    c) Prevenir el spam y el contenido no deseado en las páginas web.
    d) Clasificar las páginas web en función de su edad y autoridad.

16. En el contexto del procesamiento de grandes conjuntos de datos, se puede asegurar que:
    a) MapReduce es un enfoque para procesar datos de forma secuencial en un solo servidor para evitar problemas de concurrencia.
    b) MapReduce divide una tarea en múltiples pasos de map y reduce que se ejecutan de forma secuencial en diferentes servidores para mejorar el rendimiento y la escalabilidad.
    c) MapReduce no es adecuado para el procesamiento de datos no estructurados.
    d) MapReduce solo puede manejar pequeñas cantidades de datos y no escala bien a grandes conjuntos de datos.

17. Implementar índices invertidos en un SRI asegura:
    a) Acelerar el proceso de búsqueda al permitir búsquedas directas por contenido en lugar de por título.
    b) Reducir la cantidad de espacio de almacenamiento necesario al comprimir los datos de los documentos.
    c) Facilitar la búsqueda sobre los datos que contienen términos específicos al mantener una lista de datos para cada término único.
    d) Incrementar la seguridad de los datos almacenados al dificultar el acceso directo a la información sin el índice correcto.

18. Analizar una red permite:
    a) Detectar posibles tendencias antes de que se conviertan en tendencia.
    b) Obtener predicciones exactas de eventos futuros en mercados financieros.
    c) Evaluar la calidad del contenido de un sitio web.
    d) Encontrar nodos “sensibles” o críticos para la red.

19. En el contexto de la representación del conocimiento basada en herencia, ¿qué caracteriza a la herencia cancelable?
    a) La herencia es el resultado del razonamiento no transitivo.
    b) Las conclusiones no están determinadas y dependen del nodo de interés.
    c) Las propiedades heredadas siempre se mantienen y no pueden anularse.
    d) No existe ambigüedad en las conclusiones obtenidas.

20. Se puede afirmar que:
    a) El término "Big Data" se refiere exclusivamente al volumen de datos que una organización maneja, sin tener en cuenta la velocidad, la variedad y la veracidad de los datos.
    b) Los SRI pueden beneficiarse de MapReduce para mejorar la RI relevante.
    c) MapReduce es un modelo de procesamiento distribuido utilizado para trabajar con grandes volúmenes de datos.
    d) Uno de los desafíos en el procesamiento de Big Data es la capacidad de gestionar y analizar datos provenientes de diversas fuentes y en diferentes formatos de manera eficiente.

Trabajo de Control
Sistemas de Recuperación de Información
Fecha: 15 de julio de 2024
Temario #25

Nombre: _____ Grupo: _____

Seleccione en cada caso la(s) respuesta(s) correcta(s). Considere también no marcar ninguna opción.

1. En la RI en el contexto de Big Data, se puede asegurar que:
   a) El almacenamiento distribuido centraliza todos los datos en un único servidor para facilitar su gestión y mantenimiento.
   b) El almacenamiento distribuido ofrece ventajas significativas en términos de escalabilidad y rendimiento en comparación con el almacenamiento centralizado.
   c) El almacenamiento distribuido reparte los datos en múltiples servidores para mejorar la disponibilidad y la redundancia del sistema.
   d) El almacenamiento distribuido es una técnica obsoleta.

2. En una empresa comercial, se tiene una red donde los nodos corresponden a los empleados y las aristas representan la frecuencia con la que colaboran en las ventas. Se busca mejorar la colaboración entre los empleados para aumentar las ventas totales, por lo que la directiva debe:
   a) Utilizar la centralidad de cercanía para identificar los empleados menos cercanos a otros en la red y asignarles tareas individuales para evitar posibles conflictos y desacuerdos en el proceso de colaboración.
   b) Utilizar la centralidad de cercanía para identificar los empleados más cercanos a otros en la red y promover la colaboración entre ellos, facilitando así la comunicación y el intercambio de conocimientos para mejorar las ventas.
   c) No considerar la centralidad de cercanía de los empleados al diseñar estrategias de colaboración, ya que todos los empleados tienen el mismo acceso a la información y los recursos.
   d) No tener en cuenta la centralidad de cercanía de los empleados al diseñar estrategias de colaboración, ya que esto podría introducir complicaciones adicionales en el proceso de trabajo.

3. Sobre el algoritmo de PageRank, visto en clase, se puede afirmar que:
   a) Evalúa la importancia de un sitio web en función de la calidad y cantidad de enlaces entrantes que recibe de otros sitios web.
   b) Solo tiene en cuenta el contenido en un sitio web para determinar su relevancia en los resultados de búsqueda.
   c) Asigna una puntuación alta a los sitios web que tienen un gran número de enlaces entrantes sin tener en cuenta la calidad de esos enlaces.
   d) Asigna una puntuación baja a los sitios web que contienen muchos enlaces salientes, ya que indica una falta de relevancia.

4. En el diseño de un videojuego de roles (RPG) en el que los personajes tienen atributos como salud, fuerza y velocidad, ¿cuál de las siguientes opciones representa mejor una implementación de la representación del conocimiento orientado a objetos?
   a) Cada personaje se representa como una lista de cadenas de texto que describen sus características físicas y habilidades.
   b) Cada personaje se representa como una función que calcula sus atributos en función de su nivel y experiencia.
   c) Cada personaje se representa como una matriz de números que almacena sus valores de atributos.
   d) Cada personaje se representa como un objeto con propiedades como salud, fuerza y velocidad, y métodos para modificar y consultar estos valores.

5. Implementar índices invertidos en un SRI asegura:
   a) Acelerar el proceso de búsqueda al permitir búsquedas directas por contenido en lugar de por título.
   b) Reducir la cantidad de espacio de almacenamiento necesario al comprimir los datos de los documentos.
   c) Facilitar la búsqueda sobre los datos que contienen términos específicos al mantener una lista de datos para cada término único.
   d) Incrementar la seguridad de los datos almacenados al dificultar el acceso directo a la información sin el índice correcto.

6. La Web 2.0 se caracteriza por:
   a) La existencia de sitios web dinámicos e interactivos que permiten a los usuarios participar, comentar e interactuar tanto con los creadores de contenido como con otros usuarios.

b) El uso de distintas tecnologías para crear experiencias web más interactivas y con mayor capacidad de respuesta.

c) Las actualizaciones del contenido de los sitios web se efectúa de forma manual dentro del código HTML

d) Las plataformas ofrecen experiencias a la medida, permitiendo a los usuarios personalizar sus perfiles, recibir recomendaciones ajustadas al contenido y participar en filtrado colaborativo.

7. Se puede afirmar que:

a) El término "Big Data" se refiere exclusivamente al volumen de datos que una organización maneja, sin tener en cuenta la velocidad, la variedad y la veracidad de los datos.

b) Los SRI pueden beneficiarse de MapReduce para mejorar la RI relevante.

c) MapReduce es un modelo de procesamiento distribuido utilizado para trabajar con grandes volúmenes de datos.

d) Uno de los desafíos en el procesamiento de Big Data es la capacidad de gestionar y analizar datos provenientes de diversas fuentes y en diferentes formatos de manera eficiente.

8. Para transformar el contenido no estructurado de las páginas web en datos estructurados, el Web Scraping:

a) Extrae información basada en patrones de HTML/CSS.

b) Analiza los protocolos de red.

c) Interpreta el código JavaScript en tiempo real.

d) Convierte de forma automática imágenes a texto.

9. ¿Cuál de las siguientes opciones describe mejor la diferencia clave entre Web Crawling y Web Scraping?

a) Web Crawling se centra en la exploración y recopilación de enlaces de múltiples sitios web, mientras que Web Scraping se enfoca en la extracción específica de datos de páginas web individuales.

b) Web Crawling se realiza utilizando herramientas de automatización como Selenium WebDriver, mientras que Web Scraping se lleva a cabo mediante el análisis de HTML y CSS.

c) Web Scraping implica el análisis de la estructura y el contenido de las páginas web para extraer datos, mientras que Web Crawling se refiere a la descarga y almacenamiento de páginas web completas.

d) Web Scraping es más eficaz para rastrear e indexar contenido web para motores de búsqueda, mientras que Web Crawling se utiliza principalmente para la extracción de datos en proyectos de investigación.

10. Al realizar la optimización de contenido para SEO debe considerarse:

a) Seleccionar las palabras clave al azar.

b) Utilizar etiquetas de título y meta descripciones únicas y relevantes para cada página.

c) Incluir palabras clave de manera excesiva en el contenido para mejorar el posicionamiento.

d) Crear contenido valioso y original que satisfaga las necesidades de los usuarios.

11. Al "relajar" el concepto de clique en la detección de comunidades se intenta solucionar:

a) La imposibilidad de conectar nodos distantes.

b) La necesidad de datos externos para analizar la red.

c) La uniformidad de los nodos en términos de grado.

d) El solapamiento y la complejidad computacional.

12. Una empresa de comercio electrónico necesita mejorar su motor de búsqueda para proporcionar resultados más relevantes a sus usuarios. Actualmente, los resultados de la búsqueda no son precisos y los usuarios a menudo encuentran dificultades para encontrar productos específicos. La empresa está considerando implementar una indexación por conceptos para mejorar la relevancia de los resultados de búsqueda. ¿Qué beneficios podría provocar este cambio?

a) Permite adaptarse fácilmente a cambios en el vocabulario y la terminología utilizada en las descripciones de los productos.

b) Ayuda a identificar automáticamente relaciones entre productos, lo que puede mejorar las recomendaciones personalizadas a los usuarios.

c) Facilita la visualización de productos al agruparlos por categorías o características comunes.

d) Mejora la precisión en la búsqueda de productos relacionados, incluso cuando no coinciden exactamente con los términos de búsqueda del usuario.

13. En una red de transporte donde cada nodo es una parada de autobús y las aristas representan si existe un carro que pasa por ambos sitios, ¿qué puede mejorar el sistema de transporte?

a) No considerar la centralidad de intermediación de las estaciones de transporte público al planificar rutas y horarios, ya que no tiene impacto en las conexiones entre las paradas.

b) Utilizar el grafo inducido de los nodos con mayor valor en la centralidad de grado para aplicar la centralidad de intermediación con el propósito de reforzar las paradas con mayor tráfico.

c) Utilizar la centralidad de intermediación para identificar las paradas de transferencia clave y establecer nuevas conexiones entre diferentes líneas de transporte público.

d) No implementar sistemas de información en tiempo real para los usuarios, ya que pueden aumentar la carga de los trabajadores.

14. Los Web Crawlers se enfrentan a desafíos constantes. Dentro de ellos se encuentran:
    a) La dificultad para generar el contenido dinámico en tiempo real.
    b) La modificación del código y la estructura del sitio web.
    c) La incapacidad para interpretar correctamente el lenguaje de programación utilizado en el desarrollo de los sitios web.
    d) La falta de acceso a la base de datos del servidor web para extraer información actualizada.

15. El uso de Hadoop y MapReduce en el contexto de la RI tiene como objetivo:
    a) Automatizar el mantenimiento de sistemas de bases de datos.
    b) Facilitar el análisis en tiempo real de datos de redes sociales.
    c) Mejorar la eficiencia energética en centros de datos.
    d) Procesar y analizar grandes conjuntos de datos para la RI.

16. Analizar una red permite:
    a) Detectar posibles tendencias antes de que se conviertan en tendencia.
    b) Obtener predicciones exactas de eventos futuros en mercados financieros.
    c) Evaluar la calidad del contenido de un sitio web.
    d) Encontrar nodos "sensibles" o críticos para la red.

17. En un SRI la indexación:
    a) Consiste en asociar un identificador único a cada dato almacenado en el sistema.
    b) Mejora la experiencia del usuario.
    c) Permite la organización y la categorización de la información.
    d) Facilita la RI relevante.

18. El concepto de *rank sink* en el algoritmo de PageRank representa páginas web:
    a) Con distribuciones uniformes de la puntuación de PageRank.
    b) Que tienen una puntuación de PageRank más alta que otras debido a la manipulación de enlaces entrantes y salientes.
    c) Con una baja calidad de contenido y una cantidad insuficiente de enlaces salientes, lo que las hace menos relevantes en los resultados de búsqueda.
    d) Con un alto número de enlaces salientes que no reciben enlaces entrantes, lo que puede afectar negativamente su puntuación de PageRank.

19. Con respecto a WordNet y su contribución a las ontologías y la representación del conocimiento en los sistemas de RI, puede afirmarse que:
    a) En WordNet, una palabra está asociada a un conjunto de sinónimos (synsets), siendo estas palabras intercambiables en un contexto.
    b) La integración de WordNet en sistemas de RI permite la expansión de consultas y la mejora de la precisión de los resultados al entender mejor el significado de los términos de búsqueda a través de su contexto semántico y las relaciones entre palabras.
    c) WordNet diferencia claramente entre los significados de palabras según su uso en diferentes contextos, lo que permite aplicaciones avanzadas en desambiguación semántica más allá de los sistemas de recomendación y RI.
    d) Una limitación de WordNet en la representación del conocimiento es su enfoque en el idioma inglés, lo que plantea desafíos en la aplicación global y la interoperabilidad con sistemas de información multilingües.

20. La web actual se enfrenta a problemas como:
    a) Presencia de grandes volúmenes de datos estructurados.
    b) Presencia de una alta calidad en los datos.
    c) Presencia de datos volátiles y distribuidos.
    d) Heterogeneidad en los datos.

Trabajo de Control
Sistemas de Recuperación de Información
Fecha: 15 de julio de 2024
Temario #26

Nombre: _____ Grupo: _____

Seleccione en cada caso la(s) respuesta(s) correcta(s). Considere también no marcar ninguna opción.

1. Al "relajar" el concepto de clique en la detección de comunidades se intenta solucionar:
   a) La imposibilidad de conectar nodos distantes.
   b) La necesidad de datos externos para analizar la red.
   c) La uniformidad de los nodos en términos de grado.
   d) El solapamiento y la complejidad computacional.

2. La transición de la Web 1.0 a la Web 2.0 se caracterizó principalmente por:
   a) El aumento en la velocidad de conexión a internet, que permitió una mejor calidad de las páginas web.
   b) La disminución de la importancia de los motores de búsqueda en la navegación web.
   c) El cambio de páginas web estáticas a dinámicas, permitiendo la interacción del usuario y la generación de contenido.
   d) La reducción en el uso de HTML y CSS en el desarrollo de sitios web.

3. Para contribuir positivamente al posicionamiento orgánico de un sitio web en los motores de búsqueda se puede:
   a) Crear contenido relevante y de alta calidad que satisfaga las necesidades de información de los usuarios.
   b) Mejorar la velocidad de carga del sitio web y asegurar que sea *responsive* y fácil de usar en dispositivos móviles.
   c) Incluir una densidad alta de palabras clave para asegurar que el sitio web aparezca en tantas búsquedas como sea posible.
   d) Obtener enlaces entrantes de otros sitios web de alta autoridad y relevancia temática.

4. La computación evolutiva:
   a) No es aplicable en la RI debido a la complejidad de los algoritmos evolutivos.
   b) Solo puede manejar conjuntos de datos pequeños y no es escalable a grandes volúmenes de datos.
   c) Utiliza algoritmos para buscar soluciones óptimas en grandes espacios de búsqueda, lo que la hace adecuada para problemas de optimización en la RI.
   d) Es útil solo para problemas de clasificación de documentos y no para otras tareas de RI en general.

5. Al realizar la optimización de contenido para SEO debe considerarse:
   a) Seleccionar las palabras clave al azar.
   b) Utilizar etiquetas de título y meta descripciones únicas y relevantes para cada página.
   c) Incluir palabras clave de manera excesiva en el contenido para mejorar el posicionamiento.
   d) Crear contenido valioso y original que satisfaga las necesidades de los usuarios.

6. En un sistema donde el conocimiento está definido a partir de reglas se puede asegurar que:
   a) El nivel de especificidad de las reglas está limitado.
   b) La representación de conocimiento está basada en la lógica proposicional.
   c) El razonamiento se activa en cadena hacia delante.
   d) El orden en que se definen las reglas no altera el razonamiento del sistema.

7. En la RI en el contexto de Big Data, se puede asegurar que:
   a) El almacenamiento distribuido centraliza todos los datos en un único servidor para facilitar su gestión y mantenimiento.
   b) El almacenamiento distribuido ofrece ventajas significativas en términos de escalabilidad y rendimiento en comparación con el almacenamiento centralizado.
   c) El almacenamiento distribuido reparte los datos en múltiples servidores para mejorar la disponibilidad y la redundancia del sistema.
   d) El almacenamiento distribuido es una técnica obsoleta.

8. La Web 3 se conoce como:
   a) Internet de las cosas.
   b) Web de solo lectura.
   c) Web semántica.
   d) Web de escritura-lectura.

9. En el contexto de la RI en redes, se puede afirmar que:
   a) La detección de comunidades no es relevante para la RI en redes, ya que se centra únicamente en la estructura de la red sin considerar el contenido.

b) La detección de comunidades en una red siempre produce resultados objetivos y consistentes independientemente del algoritmo utilizado.

c) La detección de comunidades ayuda a identificar grupos de nodos altamente conectados entre sí, lo que puede ser útil para comprender la estructura y el contenido de la red.

d) La detección de comunidades solo se aplica a redes pequeñas y simples, no a redes grandes y complejas.

10. En un sistema donde el conocimiento está orientado a objetos se puede asegurar que:

a) Se enfatiza la atención a la información de la cual se extrajo el conocimiento.

b) Los marcos y las bandas son las estructuras utilizadas para la representación del modelo.

c) No es posible definir especializaciones de los objetos de la vida real dentro del sistema.

d) Las funciones de agregación dificultan poder establecer relaciones entre los objetos.

11. En el contexto del procesamiento de grandes conjuntos de datos, se puede asegurar que:

a) MapReduce es un enfoque para procesar datos de forma secuencial en un solo servidor para evitar problemas de concurrencia.

b) MapReduce divide una tarea en múltiples pasos de map y reduce que se ejecutan de forma secuencial en diferentes servidores para mejorar el rendimiento y la escalabilidad.

c) MapReduce no es adecuado para el procesamiento de datos no estructurados.

d) MapReduce solo puede manejar pequeñas cantidades de datos y no escala bien a grandes conjuntos de datos.

12. ¿Qué es un "Uniform Resource Locator (URL) Frontier" en el contexto de Web Crawling?

a) Un protocolo que define cómo se deben formatear las URLs para el crawling.

b) Una lista prioritaria de URLs que aún no han sido visitadas por el crawler.

c) Una técnica para filtrar URLs irrelevantes y mejorar la eficiencia del crawling.

d) Una base de datos que almacena URLs únicas identificadas como recursos en la web.

13. El algoritmo Hypertext Induced Topic Selection (HITS) intenta buscar nodos especiales. Estos son conocidos como:

a) Autoridades.

b) Hubs.

c) Centrales.

d) Sensibles.

14. Un SRI es capaz de:

a) Crear los índices asociados a los datos sin tener que analizar cada dato.

b) Generar índices invertidos de manera óptima sin considerar el contexto.

c) Reducir el tiempo de indexación de los datos si utiliza servidores distribuidos dentro de la red para que cada uno ejecute la indexación del mismo conjunto de datos.

d) No necesitar de ningún almacenamiento externo para alojar los índices de los datos.

15. La centralidad de intermediación de un nodo indica:

a) La cantidad de nodos vecinos directos.

b) La resistencia del nodo a fallos.

c) El número total de conexiones entrantes y salientes.

d) La frecuencia con la que un nodo actúa como puente en el camino más corto entre otros dos nodos.

16. En el diseño de un videojuego de roles (RPG) en el que los personajes tienen atributos como salud, fuerza y velocidad, ¿cuál de las siguientes opciones representa mejor una implementación de la representación del conocimiento orientado a objetos?

a) Cada personaje se representa como una lista de cadenas de texto que describen sus características físicas y habilidades.

b) Cada personaje se representa como una función que calcula sus atributos en función de su nivel y experiencia.

c) Cada personaje se representa como una matriz de números que almacena sus valores de atributos.

d) Cada personaje se representa como un objeto con propiedades como salud, fuerza y velocidad, y métodos para modificar y consultar estos valores.

17. Considerando las prácticas éticas y legales en el Web Scraping, se puede asegurar que:

a) Web Scraping requiere considerar las políticas de `robots.txt` del sitio web objetivo.

b) Es importante revisar y respetar los términos de servicio del sitio web, así como las leyes aplicables de protección de datos y derechos de autor, antes de realizar Web Scraping.

c) Web Scraping sobre datos personales sin consentimiento es generalmente aceptado si los datos se utilizan con fines de investigación.

d) La extracción de datos mediante Web Scraping siempre es legal, independientemente de las leyes locales sobre derechos de autor y privacidad de datos definidos en los sitios web.

18. El posicionamiento de un sitio web en los motores de búsqueda puede ser afectado por:

a) Utilizar técnicas de encubrimiento para mostrar contenido diferente a lo indexado por los motores de búsqueda.

b) Contenido sin valor en el sitio web.

c) Mantener una estructura de URL clara y coherente.

d) Obtener enlaces de sitios web irrelevantes y de baja calidad.

19. En un grafo una comunidad es:

a) Un conjunto de nodos que no comparten ninguna similitud estructural o funcional entre sí.

b) Un conjunto de nodos aislados.

c) Un conjunto de nodos altamente conectados que forman un subgrafo completamente independiente del resto de la red.

d) Un conjunto de nodos que están más densamente interconectados entre sí que con los nodos fuera del conjunto.

20. El concepto de *rank sink* en el algoritmo de PageRank representa páginas web:

a) Con distribuciones uniformes de la puntuación de PageRank.

b) Que tienen una puntuación de PageRank más alta que otras debido a la manipulación de enlaces entrantes y salientes.

c) Con una baja calidad de contenido y una cantidad insuficiente de enlaces salientes, lo que las hace menos relevantes en los resultados de búsqueda.

d) Con un alto número de enlaces salientes que no reciben enlaces entrantes, lo que puede afectar negativamente su puntuación de PageRank.

Trabajo de Control
Sistemas de Recuperación de Información
Fecha: 15 de julio de 2024
Temario #27

Nombre: _____ Grupo: _____

Seleccione en cada caso la(s) respuesta(s) correcta(s). Considere también no marcar ninguna opción.

1. Dentro del ecosistema de Hadoop, el HDFS se caracteriza por:
   a) El modelo de acceso y de escritura de datos en tiempo real.
   b) La tolerancia a fallos mediante la replicación de datos.
   c) El almacenamiento exclusivo para archivos de texto.
   d) La capacidad ilimitada de almacenamiento.

2. Una buena práctica de SEO para mejorar el posicionamiento de un sitio web en los motores de búsqueda es:
   a) Obtener enlaces de otros sitios web relevantes y de calidad que apunten al sitio.
   b) Copiar contenido directamente de otros sitios web populares para aumentar la cantidad de páginas indexadas.
   c) Utilizar técnicas de encubrimiento para mostrar contenido diferente a lo indexado por los motores de búsqueda.
   d) Llenar el contenido con palabras clave irrelevantes para aumentar la densidad de palabras clave.

3. Una empresa de comercio electrónico necesita mejorar su motor de búsqueda para proporcionar resultados más relevantes a sus usuarios. Actualmente, los resultados de la búsqueda no son precisos y los usuarios a menudo encuentran dificultades para encontrar productos específicos. La empresa está considerando implementar una indexación por conceptos para mejorar la relevancia de los resultados de búsqueda. ¿Qué beneficios podría provocar este cambio?
   a) Permite adaptarse fácilmente a cambios en el vocabulario y la terminología utilizada en las descripciones de los productos.
   b) Ayuda a identificar automáticamente relaciones entre productos, lo que puede mejorar las recomendaciones personalizadas a los usuarios.
   c) Facilita la visualización de productos al agruparlos por categorías o características comunes.
   d) Mejora la precisión en la búsqueda de productos relacionados, incluso cuando no coinciden exactamente con los términos de búsqueda del usuario.

4. La web actual se enfrenta a problemas como:
   a) Presencia de grandes volúmenes de datos estructurados.
   b) Presencia de una alta calidad en los datos.
   c) Presencia de datos volátiles y distribuidos.
   d) Heterogeneidad en los datos.

5. No se considera como técnica para detectar comunidades en una red:
   a) Analizar la mutualidad de los enlaces.
   b) Usar el agrupamiento jerárquico.
   c) Utilizar el algoritmo de K-Means.
   d) Encontrar cliques de vértices de grado par.

6. En el modelo de representación del conocimiento basado en herencia se puede asegurar que:
   a) La cantidad de padres que puede tener un nodo no es mayor que 4.
   b) Las conclusiones pueden ser canceladas si el grafo es ambiguo.
   c) Solo se usa en entornos referentes a la biología.
   d) El razonamiento deducido está respaldado por al menos un camino dentro del grafo.

7. Sobre el algoritmo de PageRank, visto en clase, se puede afirmar que:
   a) Evalúa la importancia de un sitio web en función de la calidad y cantidad de enlaces entrantes que recibe de otros sitios web.
   b) Solo tiene en cuenta el contenido en un sitio web para determinar su relevancia en los resultados de búsqueda.
   c) Asigna una puntuación alta a los sitios web que tienen un gran número de enlaces entrantes sin tener en cuenta la calidad de esos enlaces.
   d) Asigna una puntuación baja a los sitios web que contienen muchos enlaces salientes, ya que indica una falta de relevancia.

8. En el diseño de un videojuego de roles (RPG) en el que los personajes tienen atributos como salud, fuerza y velocidad, ¿cuál de las siguientes opciones representa mejor una implementación de la representación del conocimiento orientado a objetos?
   a) Cada personaje se representa como una lista de cadenas de texto que describen sus características físicas y habilidades.

b) Cada personaje se representa como una función que calcula sus atributos en función de su nivel y experiencia.

c) Cada personaje se representa como una matriz de números que almacena sus valores de atributos.

d) Cada personaje se representa como un objeto con propiedades como salud, fuerza y velocidad, y métodos para modificar y consultar estos valores.

9. En una red de transporte donde cada nodo es una parada de autobús y las aristas representan si existe un carro que pasa por ambos sitios, ¿qué puede mejorar el sistema de transporte?

a) No considerar la centralidad de intermediación de las estaciones de transporte público al planificar rutas y horarios, ya que no tiene impacto en las conexiones entre las paradas.

b) Utilizar el grafo inducido de los nodos con mayor valor en la centralidad de grado para aplicar la centralidad de intermediación con el propósito de reforzar las paradas con mayor tráfico.

c) Utilizar la centralidad de intermediación para identificar las paradas de transferencia clave y establecer nuevas conexiones entre diferentes líneas de transporte público.

d) No implementar sistemas de información en tiempo real para los usuarios, ya que pueden aumentar la carga de los trabajadores.

10. Para contribuir positivamente al posicionamiento orgánico de un sitio web en los motores de búsqueda se puede:

a) Crear contenido relevante y de alta calidad que satisfaga las necesidades de información de los usuarios.

b) Mejorar la velocidad de carga del sitio web y asegurar que sea *responsive* y fácil de usar en dispositivos móviles.

c) Incluir una densidad alta de palabras clave para asegurar que el sitio web aparezca en tantas búsquedas como sea posible.

d) Obtener enlaces entrantes de otros sitios web de alta autoridad y relevancia temática.

11. En la RI en el contexto de Big Data, se puede asegurar que:

a) La indexación distribuida divide los datos en múltiples fragmentos que se almacenan en varios nodos para permitir búsquedas paralelas y mejorar la escalabilidad.

b) La indexación distribuida no ofrece ventajas en términos de rendimiento y escalabilidad en comparación con la indexación centralizada.

c) La indexación distribuida es una técnica obsoleta.

d) La indexación distribuida almacena todos los datos en un solo servidor para facilitar su acceso y búsqueda.

12. Al "relajar" el concepto de clique en la detección de comunidades se intenta solucionar:

a) La imposibilidad de conectar nodos distantes.

b) La necesidad de datos externos para analizar la red.

c) La uniformidad de los nodos en términos de grado.

d) El solapamiento y la complejidad computacional.

13. Se tiene un grafo donde cada nodo es un personaje de cierto libro de cuentos y la existencia de las aristas está definida si dos personajes aparecen en el mismo cuento. Se puede asegurar que:

a) La centralidad de intermediación identifica a los personajes que actúan como conectores entre personajes de cuentos distintos.

b) La centralidad de grado es útil para identificar los personajes con más conexiones dentro de la red, lo que puede indicar su importancia en el libro.

c) La centralidad de cercanía ofrece una relación entre la cantidad de vecinos de un nodo con respecto a la longitud máxima de un camino dentro del grafo partiendo del nodo en cuestión.

d) La centralidad de cercanía indica el grado de conexión de cada personaje con el resto de los personajes de los cuentos del libro.

14. Si se tiene el conjunto de páginas interconectadas $\{A \to B, C; B \to C; C \to A\}$, entonces la página con valor más alto de PageRank es:

a) A y B.

b) A.

c) C.

d) B.

15. En el contexto de la representación del conocimiento basada en herencia, ¿qué caracteriza a la herencia cancelable?

a) La herencia es el resultado del razonamiento no transitivo.

b) Las conclusiones no están determinadas y dependen del nodo de interés.

c) Las propiedades heredadas siempre se mantienen y no pueden anularse.

d) No existe ambigüedad en las conclusiones obtenidas.

16. Referente al Web Crawler puede afirmarse que:

a) Los hipervínculos encontrados en cada sitio web que no pertenecen al dominio donde fueron encontrados se desechan, puesto que no expande el conjunto de URLs sin visitar.

b) Las páginas visitadas no se procesan nunca más.

c) No necesita de un conjunto inicial de URLs para recorrer la Web.

d) No tiene como objetivo indexar y recopilar información de diferentes sitios web.

17. El uso de Hadoop y MapReduce en el contexto de la RI tiene como objetivo:

a) Automatizar el mantenimiento de sistemas de bases de datos.

b) Facilitar el análisis en tiempo real de datos de redes sociales.

c) Mejorar la eficiencia energética en centros de datos.

d) Procesar y analizar grandes conjuntos de datos para la RI.

18. En el algoritmo de Indexación Basada en Clasificación Bloqueada (BSBI), ¿cuál es el paso final para crear un índice invertido para la colección completa de los datos?

a) Eliminar los términos duplicados de los índices.

b) Indexar cada bloque de forma independiente.

c) Fusionar los índices invertidos de cada bloque.

d) Dividir la colección de datos en bloques de tamaño fijo.

19. Analizar una red permite:

a) Detectar posibles tendencias antes de que se conviertan en tendencia.

b) Obtener predicciones exactas de eventos futuros en mercados financieros.

c) Evaluar la calidad del contenido de un sitio web.

d) Encontrar nodos "sensibles" o críticos para la red.

20. ¿Qué estrategia utilizan los Web Crawlers para asegurar un rastreo eficiente y respetuoso de los recursos de los sitios web?

a) Ignorar completamente el archivo `robots.txt` de los sitios web.

b) Visitar y rastrear todos los enlaces de una página web simultáneamente.

c) Extraer únicamente contenido multimedia para reducir la carga en los servidores web.

d) Seguir las directrices del archivo `robots.txt` y aplicar un retraso entre las solicitudes.

Trabajo de Control
Sistemas de Recuperación de Información
Fecha: 15 de julio de 2024
Temario #28

Nombre: _____ Grupo: _____

Seleccione en cada caso la(s) respuesta(s) correcta(s). Considere también no marcar ninguna opción.

1. La Web 3 se conoce como:
   a) Internet de las cosas.
   b) Web de solo lectura.
   c) Web semántica.
   d) Web de escritura-lectura.

2. Si se tiene el conjunto de páginas interconectadas $\{A \to B, C; B \to C; C \to A\}$, entonces la página con valor más alto de PageRank es:
   a) A y B.
   b) A.
   c) C.
   d) B.

3. Si una red cumple la propiedad de ser un grafo de mundo pequeño, entonces se conoce que:
   a) El número de componentes fuertemente conexas está relacionado con la cantidad de grafos K_n presentes.
   b) La longitud media del camino entre todo par de vértices es pequeña.
   c) El grafo es un anillo regular de grado 5.
   d) La red posee pocos vértices.

4. En el modelo de representación del conocimiento basado en herencia se puede asegurar que:
   a) La cantidad de padres que puede tener un nodo no es mayor que 4.
   b) Las conclusiones pueden ser canceladas si el grafo es ambiguo.
   c) Solo se usa en entornos referentes a la biología.
   d) El razonamiento deducido está respaldado por al menos un camino dentro del grafo.

5. En un sistema donde el conocimiento está orientado a objetos se puede asegurar que:
   a) Se enfatiza la atención a la información de la cual se extrajo el conocimiento.
   b) Los marcos y las bandas son las estructuras utilizadas para la representación del modelo.
   c) No es posible definir especializaciones de los objetos de la vida real dentro del sistema.
   d) Las funciones de agregación dificultan poder establecer relaciones entre los objetos.

6. Referente al Web Crawler puede afirmarse que:
   a) Los hipervínculos encontrados en cada sitio web que no pertenecen al dominio donde fueron encontrados se desechan, puesto que no expande el conjunto de URLs sin visitar.
   b) Las páginas visitadas no se procesan nunca más.
   c) No necesita de un conjunto inicial de URLs para recorrer la Web.
   d) No tiene como objetivo indexar y recopilar información de diferentes sitios web.

7. En un sistema donde el conocimiento está definido a partir de reglas se puede asegurar que:
   a) El nivel de especificidad de las reglas está limitado.
   b) La representación de conocimiento está basada en la lógica proposicional.
   c) El razonamiento se activa en cadena hacia delante.
   d) El orden en que se definen las reglas no altera el razonamiento del sistema.

8. Para transformar el contenido no estructurado de las páginas web en datos estructurados, el Web Scraping:
   a) Extrae información basada en patrones de HTML/CSS.
   b) Analiza los protocolos de red.
   c) Interpreta el código JavaScript en tiempo real.
   d) Convierte de forma automática imágenes a texto.

9. La computación evolutiva:
   a) No es aplicable en la RI debido a la complejidad de los algoritmos evolutivos.
   b) Solo puede manejar conjuntos de datos pequeños y no es escalable a grandes volúmenes de datos.
   c) Utiliza algoritmos para buscar soluciones óptimas en grandes espacios de búsqueda, lo que la hace adecuada para problemas de optimización en la RI.
   d) Es útil solo para problemas de clasificación de documentos y no para otras tareas de RI en general.

10. No se considera como técnica para detectar comunidades en una red:
    a) Analizar la mutualidad de los enlaces.
    b) Usar el agrupamiento jerárquico.

c) Utilizar el algoritmo de K-Means.

d) Encontrar cliques de vértices de grado par.

11. En un SRI la indexación:

a) Consiste en asociar un identificador único a cada dato almacenado en el sistema.

b) Mejora la experiencia del usuario.

c) Permite la organización y la categorización de la información.

d) Facilita la RI relevante.

12. La afirmación que mejor describe la política de amabilidad en los Web Crawlers es:

a) Los crawlers se diseñan para acceder a sitios web sin restricciones y extraer datos de manera agresiva para su indexación.

b) La política de amabilidad de los Web Crawlers dicta que los crawlers deben priorizar ciertos tipos de contenido sobre otros, ignorando completamente ciertas páginas web.

c) La política de amabilidad establece pautas y reglas sobre cómo los crawlers deben interactuar con los sitios web para minimizar la carga del servidor y respetar las directivas de los administradores del sitio.

d) Los Web Crawlers son libres de recopilar datos de cualquier sitio web sin restricciones, independientemente de la cantidad de tráfico que generen.

13. Considerando las dimensiones y desafíos inherentes a Big Data puede afirmarse, tomando en cuenta las características clave y las implicaciones para su procesamiento y análisis, que:

a) Big Data se caracteriza principalmente por su pequeño volumen y uniformidad, permitiendo un procesamiento eficiente con mínimas adaptaciones de las herramientas de análisis de datos tradicionales.

b) Big Data no desafía la capacidad de las herramientas tradicionales de procesamiento de datos para capturar, almacenar, gestionar y analizar efectivamente la información, dada la evolución constante de las capacidades computacionales y algoritmos de optimización.

c) Aunque Big Data puede incluir datos estructurados, su naturaleza se expande al incorporar grandes cantidades de datos no estructurados y semiestructurados, lo que exige el uso de tecnologías especializadas en almacenamiento y procesamiento como Hadoop y sistemas de bases de datos NoSQL.

d) Además de su complejidad y diversidad, Big Data introduce desafíos significativos en términos de veracidad y variabilidad al requerir métodos avanzados de limpieza y validación de datos para asegurar la integridad del análisis.

14. Sobre el algoritmo de PageRank, visto en clase, se puede afirmar que:

a) Evalúa la importancia de un sitio web en función de la calidad y cantidad de enlaces entrantes que recibe de otros sitios web.

b) Solo tiene en cuenta el contenido en un sitio web para determinar su relevancia en los resultados de búsqueda.

c) Asigna una puntuación alta a los sitios web que tienen un gran número de enlaces entrantes sin tener en cuenta la calidad de esos enlaces.

d) Asigna una puntuación baja a los sitios web que contienen muchos enlaces salientes, ya que indica una falta de relevancia.

15. Una buena práctica de SEO para mejorar el posicionamiento de un sitio web en los motores de búsqueda es:

a) Obtener enlaces de otros sitios web relevantes y de calidad que apunten al sitio.

b) Copiar contenido directamente de otros sitios web populares para aumentar la cantidad de páginas indexadas.

c) Utilizar técnicas de encubrimiento para mostrar contenido diferente a lo indexado por los motores de búsqueda.

d) Llenar el contenido con palabras clave irrelevantes para aumentar la densidad de palabras clave.

16. En la RI en el contexto de Big Data, se puede asegurar que:

a) El almacenamiento distribuido centraliza todos los datos en un único servidor para facilitar su gestión y mantenimiento.

b) El almacenamiento distribuido ofrece ventajas significativas en términos de escalabilidad y rendimiento en comparación con el almacenamiento centralizado.

c) El almacenamiento distribuido reparte los datos en múltiples servidores para mejorar la disponibilidad y la redundancia del sistema.

d) El almacenamiento distribuido es una técnica obsoleta.

17. En el algoritmo de Indexación Basada en Clasificación Bloqueada (BSBI), ¿cuál es el paso final para crear un índice invertido para la colección completa de los datos?

a) Eliminar los términos duplicados de los índices.

b) Indexar cada bloque de forma independiente.

c) Fusionar los índices invertidos de cada bloque.

d) Dividir la colección de datos en bloques de tamaño fijo.

18. En una empresa comercial, se tiene una red donde los nodos corresponden a los empleados y las aristas representan la frecuencia con la que colaboran en las ventas. Se busca mejorar la colaboración entre los empleados para aumentar las ventas totales, por lo que la directiva debe:

    a) Utilizar la centralidad de cercanía para identificar los empleados menos cercanos a otros en la red y asignarles tareas individuales para evitar posibles conflictos y desacuerdos en el proceso de colaboración.

    b) Utilizar la centralidad de cercanía para identificar los empleados más cercanos a otros en la red y promover la colaboración entre ellos, facilitando así la comunicación y el intercambio de conocimientos para mejorar las ventas.

    c) No considerar la centralidad de cercanía de los empleados al diseñar estrategias de colaboración, ya que todos los empleados tienen el mismo acceso a la información y los recursos.

    d) No tener en cuenta la centralidad de cercanía de los empleados al diseñar estrategias de colaboración, ya que esto podría introducir complicaciones adicionales en el proceso de trabajo.

19. Para contribuir positivamente al posicionamiento orgánico de un sitio web en los motores de búsqueda se puede:

    a) Crear contenido relevante y de alta calidad que satisfaga las necesidades de información de los usuarios.

    b) Mejorar la velocidad de carga del sitio web y asegurar que sea *responsive* y fácil de usar en dispositivos móviles.

    c) Incluir una densidad alta de palabras clave para asegurar que el sitio web aparezca en tantas búsquedas como sea posible.

    d) Obtener enlaces entrantes de otros sitios web de alta autoridad y relevancia temática.

20. El componente responsable de la gestión de recursos y planificación de tareas en Hadoop es:

    a) MapReduce.

    b) Hadoop Common.

    c) HDFS.

    d) YARN.

Trabajo de Control
Sistemas de Recuperación de Información
Fecha: 15 de julio de 2024
Temario #29

Nombre: _____ Grupo: _____

Seleccione en cada caso la(s) respuesta(s) correcta(s). Considere también no marcar ninguna opción.

1. En un grafo una comunidad es:
   a) Un conjunto de nodos que no comparten ninguna similitud estructural o funcional entre sí.
   b) Un conjunto de nodos aislados.
   c) Un conjunto de nodos altamente conectados que forman un subgrafo completamente independiente del resto de la red.
   d) Un conjunto de nodos que están más densamente interconectados entre sí que con los nodos fuera del conjunto.

2. El algoritmo Hypertext Induced Topic Selection (HITS) intenta buscar nodos especiales. Estos son conocidos como:
   a) Autoridades.
   b) Hubs.
   c) Centrales.
   d) Sensibles.

3. La política de ordenación de URLs en los Web Crawlers tiene como aspecto fundamental:
   a) Limitar el acceso de los crawlers a ciertas secciones de un sitio web, evitando el rastreo de URLs consideradas menos importantes o sensibles.
   b) Definir la estructura de la URL de destino, asegurando que estén ordenadas alfabéticamente para facilitar la navegación y la indexación.
   c) Establecer la prioridad de rastreo de las URLs, determinando el orden en que los crawlers visitan y procesan cada página web.
   d) Determinar la forma en que los crawlers asignan un valor de relevancia a cada URL para clasificarlas en el índice de búsqueda.

4. Una buena práctica de SEO para mejorar el posicionamiento de un sitio web en los motores de búsqueda es:
   a) Obtener enlaces de otros sitios web relevantes y de calidad que apunten al sitio.
   b) Copiar contenido directamente de otros sitios web populares para aumentar la cantidad de páginas indexadas.
   c) Utilizar técnicas de encubrimiento para mostrar contenido diferente a lo indexado por los motores de búsqueda.
   d) Llenar el contenido con palabras clave irrelevantes para aumentar la densidad de palabras clave.

5. El posicionamiento de un sitio web en los motores de búsqueda puede ser afectado por:
   a) Utilizar técnicas de encubrimiento para mostrar contenido diferente a lo indexado por los motores de búsqueda.
   b) Contenido sin valor en el sitio web.
   c) Mantener una estructura de URL clara y coherente.
   d) Obtener enlaces de sitios web irrelevantes y de baja calidad.

6. La computación evolutiva:
   a) No es aplicable en la RI debido a la complejidad de los algoritmos evolutivos.
   b) Solo puede manejar conjuntos de datos pequeños y no es escalable a grandes volúmenes de datos.
   c) Utiliza algoritmos para buscar soluciones óptimas en grandes espacios de búsqueda, lo que la hace adecuada para problemas de optimización en la RI.
   d) Es útil solo para problemas de clasificación de documentos y no para otras tareas de RI en general.

7. ¿Cuál de las siguientes opciones describe mejor la diferencia clave entre Web Crawling y Web Scraping?
   a) Web Crawling se centra en la exploración y recopilación de enlaces de múltiples sitios web, mientras que Web Scraping se enfoca en la extracción específica de datos de páginas web individuales.
   b) Web Crawling se realiza utilizando herramientas de automatización como Selenium WebDriver, mientras que Web Scraping se lleva a cabo mediante el análisis de HTML y CSS.
   c) Web Scraping implica el análisis de la estructura y el contenido de las páginas web para extraer datos, mientras que Web Crawling se refiere a la descarga y almacenamiento de páginas web completas.
   d) Web Scraping es más eficaz para rastrear e indexar contenido web para motores de búsqueda, mientras que Web Crawling se utiliza principalmente para la extracción de datos en proyectos de investigación.

8. La afirmación que mejor describe la política de amabilidad en los Web Crawlers es:

a) Los crawlers se diseñan para acceder a sitios web sin restricciones y extraer datos de manera agresiva para su indexación.

b) La política de amabilidad de los Web Crawlers dicta que los crawlers deben priorizar ciertos tipos de contenido sobre otros, ignorando completamente ciertas páginas web.

c) La política de amabilidad establece pautas y reglas sobre cómo los crawlers deben interactuar con los sitios web para minimizar la carga del servidor y respetar las directivas de los administradores del sitio.

d) Los Web Crawlers son libres de recopilar datos de cualquier sitio web sin restricciones, independientemente de la cantidad de tráfico que generen.

9. En la RI en el contexto de Big Data, se puede asegurar que:

a) El almacenamiento distribuido centraliza todos los datos en un único servidor para facilitar su gestión y mantenimiento.

b) El almacenamiento distribuido ofrece ventajas significativas en términos de escalabilidad y rendimiento en comparación con el almacenamiento centralizado.

c) El almacenamiento distribuido reparte los datos en múltiples servidores para mejorar la disponibilidad y la redundancia del sistema.

d) El almacenamiento distribuido es una técnica obsoleta.

10. En un sistema donde el conocimiento está orientado a objetos se puede asegurar que:

a) Se enfatiza la atención a la información de la cual se extrajo el conocimiento.

b) Los marcos y las bandas son las estructuras utilizadas para la representación del modelo.

c) No es posible definir especializaciones de los objetos de la vida real dentro del sistema.

d) Las funciones de agregación dificultan poder establecer relaciones entre los objetos.

11. La premisa básica del algoritmo de PageRank para clasificar páginas web en los resultados de búsqueda es:

a) La antigüedad de la página web es el principal factor para determinar su clasificación.

b) El contenido y la relevancia de las palabras clave en la página web determinan su posición.

c) Los enlaces entrantes a una página web desde otras páginas contribuyen a su importancia y clasificación.

d) La cantidad de visitas que recibe una página web determina su clasificación.

12. Implementar índices invertidos en un SRI asegura:

a) Acelerar el proceso de búsqueda al permitir búsquedas directas por contenido en lugar de por título.

b) Reducir la cantidad de espacio de almacenamiento necesario al comprimir los datos de los documentos.

c) Facilitar la búsqueda sobre los datos que contienen términos específicos al mantener una lista de datos para cada término único.

d) Incrementar la seguridad de los datos almacenados al dificultar el acceso directo a la información sin el índice correcto.

13. Considerando las prácticas éticas y legales en el Web Scraping, se puede asegurar que:

a) Web Scraping requiere considerar las políticas de `robots.txt` del sitio web objetivo.

b) Es importante revisar y respetar los términos de servicio del sitio web, así como las leyes aplicables de protección de datos y derechos de autor, antes de realizar Web Scraping.

c) Web Scraping sobre datos personales sin consentimiento es generalmente aceptado si los datos se utilizan con fines de investigación.

d) La extracción de datos mediante Web Scraping siempre es legal, independientemente de las leyes locales sobre derechos de autor y privacidad de datos definidos en los sitios web.

14. Se puede afirmar que:

a) No es posible identificar subgrupos dentro de una red utilizando análisis de redes.

b) La cantidad de conexiones de un nodo siempre indica su influencia en la red.

c) El tamaño de una red es siempre indicativo de su efectividad en la transmisión de información.

d) Todas las relaciones en una red tienen la misma importancia para el análisis.

15. Los algoritmos para detectar comunidades en una red intentan:

a) Buscar conjuntos donde cada nodo de un mismo conjunto tenga características similares al resto de los nodos del conjunto.

b) Seleccionar aleatoriamente nodos de alto grado y sus vecinos.

c) Buscar subgrafos tal que no incluyan nodos cuya ausencia desconecte al subgrafo.

d) Encontrar grupos donde los nodos pertenecientes a los mismos grupos son cercanos bajo cierta métrica y lejanos con respecto a los nodos de otros grupos.

16. Se tiene un grafo donde cada nodo es un personaje de cierto libro de cuentos y la existencia de las aristas está definida si dos personajes aparecen en el mismo cuento. Se puede asegurar que:

a) La centralidad de intermediación identifica a los personajes que actúan como conectores entre personajes de cuentos distintos.

b) La centralidad de grado es útil para identificar los personajes con más conexiones dentro de la red, lo que puede indicar su importancia en el libro.

c) La centralidad de cercanía ofrece una relación entre la cantidad de vecinos de un nodo con respecto a la longitud máxima de un camino dentro del grafo partiendo del nodo en cuestión.

d) La centralidad de cercanía indica el grado de conexión de cada personaje con el resto de los personajes de los cuentos del libro.

17. Un SRI es capaz de:

a) Crear los índices asociados a los datos sin tener que analizar cada dato.

b) Generar índices invertidos de manera óptima sin considerar el contexto.

c) Reducir el tiempo de indexación de los datos si utiliza servidores distribuidos dentro de la red para que cada uno ejecute la indexación del mismo conjunto de datos.

d) No necesitar de ningún almacenamiento externo para alojar los índices de los datos.

18. En un sistema donde el conocimiento está definido a partir de reglas se puede asegurar que:

a) El nivel de especificidad de las reglas está limitado.

b) La representación de conocimiento está basada en la lógica proposicional.

c) El razonamiento se activa en cadena hacia delante.

d) El orden en que se definen las reglas no altera el razonamiento del sistema.

19. En la RI en el contexto de Big Data, se puede asegurar que:

a) La indexación distribuida divide los datos en múltiples fragmentos que se almacenan en varios nodos para permitir búsquedas paralelas y mejorar la escalabilidad.

b) La indexación distribuida no ofrece ventajas en términos de rendimiento y escalabilidad en comparación con la indexación centralizada.

c) La indexación distribuida es una técnica obsoleta.

d) La indexación distribuida almacena todos los datos en un solo servidor para facilitar su acceso y búsqueda.

20. El concepto de *rank sink* en el algoritmo de PageRank representa páginas web:

a) Con distribuciones uniformes de la puntuación de PageRank.

b) Que tienen una puntuación de PageRank más alta que otras debido a la manipulación de enlaces entrantes y salientes.

c) Con una baja calidad de contenido y una cantidad insuficiente de enlaces salientes, lo que las hace menos relevantes en los resultados de búsqueda.

d) Con un alto número de enlaces salientes que no reciben enlaces entrantes, lo que puede afectar negativamente su puntuación de PageRank.

Trabajo de Control
Sistemas de Recuperación de Información
Fecha: 15 de julio de 2024
Temario #30

Nombre: _____ Grupo: _____

Seleccione en cada caso la(s) respuesta(s) correcta(s). Considere también no marcar ninguna opción.

1. Los algoritmos para detectar comunidades en una red intentan:
   a) Buscar conjuntos donde cada nodo de un mismo conjunto tenga características similares al resto de los nodos del conjunto.
   b) Seleccionar aleatoriamente nodos de alto grado y sus vecinos.
   c) Buscar subgrafos tal que no incluyan nodos cuya ausencia desconecte al subgrafo.
   d) Encontrar grupos donde los nodos pertenecientes a los mismos grupos son cercanos bajo cierta métrica y lejanos con respecto a los nodos de otros grupos.

2. La premisa básica del algoritmo de PageRank para clasificar páginas web en los resultados de búsqueda es:
   a) La antigüedad de la página web es el principal factor para determinar su clasificación.
   b) El contenido y la relevancia de las palabras clave en la página web determinan su posición.
   c) Los enlaces entrantes a una página web desde otras páginas contribuyen a su importancia y clasificación.
   d) La cantidad de visitas que recibe una página web determina su clasificación.

3. Una empresa de comercio electrónico necesita mejorar su motor de búsqueda para proporcionar resultados más relevantes a sus usuarios. Actualmente, los resultados de la búsqueda no son precisos y los usuarios a menudo encuentran dificultades para encontrar productos específicos. La empresa está considerando implementar una indexación por conceptos para mejorar la relevancia de los resultados de búsqueda. ¿Qué beneficios podría provocar este cambio?
   a) Permite adaptarse fácilmente a cambios en el vocabulario y la terminología utilizada en las descripciones de los productos.
   b) Ayuda a identificar automáticamente relaciones entre productos, lo que puede mejorar las recomendaciones personalizadas a los usuarios.
   c) Facilita la visualización de productos al agruparlos por categorías o características comunes.
   d) Mejora la precisión en la búsqueda de productos relacionados, incluso cuando no coinciden exactamente con los términos de búsqueda del usuario.

4. Para la RI, el análisis de las redes puede:
   a) Ayudar a identificar grupos de interés.
   b) Indicar la importancia de una entidad en la transmisión de la información.
   c) Ayudar a comprender la conectividad y la accesibilidad entre las entidades.
   d) Revelar patrones de influencia dentro de una comunidad.

5. Un sistema cuenta con la siguiente información:
   - Juan pertenece a la marina.
   - Juan es capellán (sacerdote encargado de una tarea específica fuera de la parroquia).
   - Los infantes de la marina suelen ser bebedores de cerveza.
   - Un capellán no suele ser bebedor de cerveza.
   - Un bebedor de cerveza suele tener sobrepeso.
   - Por lo general, un infante de la marina está en buena condición física.

   Si se representa la información en una red de herencia se puede concluir que:
   a) No se puede asegurar que Juan sea capellán.
   b) No se puede asegurar que Juan esté en buena condición física.
   c) Juan tiene sobrepeso.
   d) Existen dos razonamientos cancelables.

6. Un SRI es capaz de:
   a) Crear los índices asociados a los datos sin tener que analizar cada dato.
   b) Generar índices invertidos de manera óptima sin considerar el contexto.
   c) Reducir el tiempo de indexación de los datos si utiliza servidores distribuidos dentro de la red para que cada uno ejecute la indexación del mismo conjunto de datos.
   d) No necesitar de ningún almacenamiento externo para alojar los índices de los datos.

7. El concepto de *rank sink* en el algoritmo de PageRank representa páginas web:

a) Con distribuciones uniformes de la puntuación de PageRank.

b) Que tienen una puntuación de PageRank más alta que otras debido a la manipulación de enlaces entrantes y salientes.

c) Con una baja calidad de contenido y una cantidad insuficiente de enlaces salientes, lo que las hace menos relevantes en los resultados de búsqueda.

d) Con un alto número de enlaces salientes que no reciben enlaces entrantes, lo que puede afectar negativamente su puntuación de PageRank.

8. En un SRI la indexación:

a) Consiste en asociar un identificador único a cada dato almacenado en el sistema.

b) Mejora la experiencia del usuario.

c) Permite la organización y la categorización de la información.

d) Facilita la RI relevante.

9. En la RI en el contexto de Big Data, se puede asegurar que:

a) La indexación distribuida divide los datos en múltiples fragmentos que se almacenan en varios nodos para permitir búsquedas paralelas y mejorar la escalabilidad.

b) La indexación distribuida no ofrece ventajas en términos de rendimiento y escalabilidad en comparación con la indexación centralizada.

c) La indexación distribuida es una técnica obsoleta.

d) La indexación distribuida almacena todos los datos en un solo servidor para facilitar su acceso y búsqueda.

10. Una plataforma de comercio electrónico desea mejorar la experiencia del usuario al permitir una navegación más personalizada y contextualizada. Actualmente los usuarios tienen dificultades para encontrar productos relevantes debido a la gran cantidad de opciones disponibles. La empresa está interesada en implementar características de la Web 2.5 y la Web Semántica para abordar este problema. ¿Qué características podrían ayudar para ofrecer una navegación más personalizada y contextualizada?

a) La implementación de ontologías y metadatos para enriquecer la descripción de productos y mejorar la precisión de las recomendaciones.

b) La optimización de la velocidad de carga del sitio web para mejorar la experiencia del usuario y reducir el abandono del carrito de compra.

c) La integración de redes sociales para permitir la recomendación de productos basada en las preferencias de amigos y contactos.

d) La incorporación de los productos en tendencia en el mercado.

11. La Web 1.0 se caracteriza por:

a) Las actualizaciones del contenido de los sitios web se efectúa de forma manual dentro del código HTML

b) Los sitios web se centran en brindar información en lugar de facilitar la colaboración o participación de los usuarios.

c) Uso de comunidades virtuales para popularizar los sitios web de noticias.

d) Los propietarios de los sitios web proporcionan contenido de forma periódica.

12. Se tiene una red de ingredientes donde cada uno representa un nodo y las aristas simbolizan que los ingredientes forman parte de una misma receta. Se busca mejorar la experiencia culinaria mediante la elaboración de combinaciones de ingredientes más interesantes y creativas, para lo cual se debe:

a) No considerar la centralidad de grado de los ingredientes, ya que todas las combinaciones de ingredientes son igualmente válidas.

b) Utilizar la centralidad de grado para identificar los ingredientes menos conectados en la red y tomarlos en cuenta para su inclusión en las combinaciones.

c) No tener en cuenta la centralidad de grado de los ingredientes para no darle mayor importancia a los ingredientes más comunes.

d) Utilizar la centralidad de grado para identificar los ingredientes más populares en la red y crear combinaciones que incluyan una variedad de ingredientes menos comunes.

13. El algoritmo de PageRank converge si:

a) La norma de la diferencia entre los vectores es menor a un umbral predefinido.

b) Finaliza la ejecución cuando el número de iteraciones excede un máximo de iteraciones previamente definido.

c) Se define un factor de normalización en la fórmula de la función.

d) El algoritmo no se implementa de forma iterativa.

14. El uso de Hadoop y MapReduce en el contexto de la RI tiene como objetivo:

a) Automatizar el mantenimiento de sistemas de bases de datos.

b) Facilitar el análisis en tiempo real de datos de redes sociales.

c) Mejorar la eficiencia energética en centros de datos.

d) Procesar y analizar grandes conjuntos de datos para la RI.

15. En un sistema donde el conocimiento está definido a partir de reglas se puede asegurar que:

a) El nivel de especificidad de las reglas está limitado.

b) La representación de conocimiento está basada en la lógica proposicional.

c) El razonamiento se activa en cadena hacia delante.

d) El orden en que se definen las reglas no altera el razonamiento del sistema.

16. El posicionamiento de un sitio web en los motores de búsqueda puede ser afectado por:

a) Utilizar técnicas de encubrimiento para mostrar contenido diferente a lo indexado por los motores de búsqueda.

b) Contenido sin valor en el sitio web.

c) Mantener una estructura de URL clara y coherente.

d) Obtener enlaces de sitios web irrelevantes y de baja calidad.

17. Dentro del ecosistema de Hadoop, el HDFS se caracteriza por:

a) El modelo de acceso y de escritura de datos en tiempo real.

b) La tolerancia a fallos mediante la replicación de datos.

c) El almacenamiento exclusivo para archivos de texto.

d) La capacidad ilimitada de almacenamiento.

18. Para transformar el contenido no estructurado de las páginas web en datos estructurados, el Web Scraping:

a) Extrae información basada en patrones de HTML/CSS.

b) Analiza los protocolos de red.

c) Interpreta el código JavaScript en tiempo real.

d) Convierte de forma automática imágenes a texto.

19. ¿Cuál de las siguientes opciones describe mejor la diferencia clave entre Web Crawling y Web Scraping?

a) Web Crawling se centra en la exploración y recopilación de enlaces de múltiples sitios web, mientras que Web Scraping se enfoca en la extracción específica de datos de páginas web individuales.

b) Web Crawling se realiza utilizando herramientas de automatización como Selenium WebDriver, mientras que Web Scraping se lleva a cabo mediante el análisis de HTML y CSS.

c) Web Scraping implica el análisis de la estructura y el contenido de las páginas web para extraer datos, mientras que Web Crawling se refiere a la descarga y almacenamiento de páginas web completas.

d) Web Scraping es más eficaz para rastrear e indexar contenido web para motores de búsqueda, mientras que Web Crawling se utiliza principalmente para la extracción de datos en proyectos de investigación.

20. La afirmación que mejor describe la política de amabilidad en los Web Crawlers es:

a) Los crawlers se diseñan para acceder a sitios web sin restricciones y extraer datos de manera agresiva para su indexación.

b) La política de amabilidad de los Web Crawlers dicta que los crawlers deben priorizar ciertos tipos de contenido sobre otros, ignorando completamente ciertas páginas web.

c) La política de amabilidad establece pautas y reglas sobre cómo los crawlers deben interactuar con los sitios web para minimizar la carga del servidor y respetar las directivas de los administradores del sitio.

d) Los Web Crawlers son libres de recopilar datos de cualquier sitio web sin restricciones, independientemente de la cantidad de tráfico que generen.

Trabajo de Control
Sistemas de Recuperación de Información
Fecha: 15 de julio de 2024
Temario #31

Nombre: _____ Grupo: _____

Seleccione en cada caso la(s) respuesta(s) correcta(s). Considere también no marcar ninguna opción.

1. ¿Qué algoritmos permiten obtener información de una red?
   a) Índices de centralidad.
   b) Detección de comunidades.
   c) Hypertext Induced Topic Selection (HITS).
   d) PageRank.

2. Si se tiene el conjunto de páginas interconectadas $\{A \to B, C; B \to C; C \to A\}$, entonces la página con valor más alto de PageRank es:
   a) A y B.
   b) A.
   c) C.
   d) B.

3. Un sistema cuenta con la siguiente información:
   - Juan pertenece a la marina.
   - Juan es capellán (sacerdote encargado de una tarea específica fuera de la parroquia).
   - Los infantes de la marina suelen ser bebedores de cerveza.
   - Un capellán no suele ser bebedor de cerveza.
   - Un bebedor de cerveza suele tener sobrepeso.
   - Por lo general, un infante de la marina está en buena condición física.

   Si se representa la información en una red de herencia se puede concluir que:
   a) No se puede asegurar que Juan sea capellán.
   b) No se puede asegurar que Juan esté en buena condición física.
   c) Juan tiene sobrepeso.
   d) Existen dos razonamientos cancelables.

4. En un sistema de control de tráfico urbano basado en reglas, ¿cuál de las siguientes reglas sería más efectiva para manejar situaciones de congestión en una intersección?
   a) Si hay pocos vehículos en la intersección, reducir el tiempo de los semáforos en verde.
   b) Si hay muchos vehículos en la intersección, aumentar el tiempo de los semáforos en verde.
   c) Si hay un vehículo de emergencia en la intersección, detener todos los demás vehículos.
   d) Si hay muchos peatones cruzando la intersección, reducir el tiempo de los semáforos en rojo.

5. Analizar una red permite:
   a) Detectar posibles tendencias antes de que se conviertan en tendencia.
   b) Obtener predicciones exactas de eventos futuros en mercados financieros.
   c) Evaluar la calidad del contenido de un sitio web.
   d) Encontrar nodos "sensibles" o críticos para la red.

6. La integración de Hadoop y MapReduce en la RI trae como ventaja:
   a) La eliminación de la necesidad de sistemas de bases de datos.
   b) La posibilidad de la extracción de información relevante y la generación de resultados significativos de grandes conjuntos de datos.
   c) La garantía de la privacidad absoluta de los datos procesados.
   d) La reducción de los costos operativos a cero.

7. La premisa básica del algoritmo de PageRank para clasificar páginas web en los resultados de búsqueda es:
   a) La antigüedad de la página web es el principal factor para determinar su clasificación.
   b) El contenido y la relevancia de las palabras clave en la página web determinan su posición.
   c) Los enlaces entrantes a una página web desde otras páginas contribuyen a su importancia y clasificación.
   d) La cantidad de visitas que recibe una página web determina su clasificación.

8. Si una red cumple la propiedad de ser un grafo de mundo pequeño, entonces se conoce que:
   a) El número de componentes fuertemente conexas está relacionado con la cantidad de grafos K_n presentes.
   b) La longitud media del camino entre todo par de vértices es pequeña.
   c) El grafo es un anillo regular de grado 5.
   d) La red posee pocos vértices.

9. La web actual se enfrenta a problemas como:
   a) Presencia de grandes volúmenes de datos estructurados.
   b) Presencia de una alta calidad en los datos.
   c) Presencia de datos volátiles y distribuidos.
   d) Heterogeneidad en los datos.

10. En la RI en el contexto de Big Data, se puede asegurar que:
   a) La indexación distribuida divide los datos en múltiples fragmentos que se almacenan en varios nodos para permitir búsquedas paralelas y mejorar la escalabilidad.
   b) La indexación distribuida no ofrece ventajas en términos de rendimiento y escalabilidad en comparación con la indexación centralizada.
   c) La indexación distribuida es una técnica obsoleta.
   d) La indexación distribuida almacena todos los datos en un solo servidor para facilitar su acceso y búsqueda.

11. Referente al Web Crawler puede afirmarse que:
   a) Los hipervínculos encontrados en cada sitio web que no pertenecen al dominio donde fueron encontrados se desechan, puesto que no expande el conjunto de URLs sin visitar.
   b) Las páginas visitadas no se procesan nunca más.
   c) No necesita de un conjunto inicial de URLs para recorrer la Web.
   d) No tiene como objetivo indexar y recopilar información de diferentes sitios web.

12. El posicionamiento de un sitio web en los motores de búsqueda puede ser afectado por:
   a) Utilizar técnicas de encubrimiento para mostrar contenido diferente a lo indexado por los motores de búsqueda.
   b) Contenido sin valor en el sitio web.
   c) Mantener una estructura de URL clara y coherente.
   d) Obtener enlaces de sitios web irrelevantes y de baja calidad.

13. La Web 2.0 se caracteriza por:
   a) La existencia de sitios web dinámicos e interactivos que permiten a los usuarios participar, comentar e interactuar tanto con los creadores de contenido como con otros usuarios.
   b) El uso de distintas tecnologías para crear experiencias web más interactivas y con mayor capacidad de respuesta.
   c) Las actualizaciones del contenido de los sitios web se efectúa de forma manual dentro del código HTML
   d) Las plataformas ofrecen experiencias a la medida, permitiendo a los usuarios personalizar sus perfiles, recibir recomendaciones ajustadas al contenido y participar en filtrado colaborativo.

14. Para la RI, el análisis de las redes puede:
   a) Ayudar a identificar grupos de interés.
   b) Indicar la importancia de una entidad en la transmisión de la información.
   c) Ayudar a comprender la conectividad y la accesibilidad entre las entidades.
   d) Revelar patrones de influencia dentro de una comunidad.

15. Para transformar el contenido no estructurado de las páginas web en datos estructurados, el Web Scraping:
   a) Extrae información basada en patrones de HTML/CSS.
   b) Analiza los protocolos de red.
   c) Interpreta el código JavaScript en tiempo real.
   d) Convierte de forma automática imágenes a texto.

16. En el diseño de un videojuego de roles (RPG) en el que los personajes tienen atributos como salud, fuerza y velocidad, ¿cuál de las siguientes opciones representa mejor una implementación de la representación del conocimiento orientado a objetos?
   a) Cada personaje se representa como una lista de cadenas de texto que describen sus características físicas y habilidades.
   b) Cada personaje se representa como una función que calcula sus atributos en función de su nivel y experiencia.
   c) Cada personaje se representa como una matriz de números que almacena sus valores de atributos.
   d) Cada personaje se representa como un objeto con propiedades como salud, fuerza y velocidad, y métodos para modificar y consultar estos valores.

17. El propósito de la política de revisitado en los Web Crawlers es:
   a) Determinar la frecuencia con la que los crawlers deben volver a visitar una URL específica para mantener la información actualizada en el índice de búsqueda.
   b) Establecer reglas sobre el tiempo máximo que los crawlers pueden pasar en un sitio web durante cada visita para evitar sobrecargar los servidores.
   c) Limitar el acceso de los crawlers a ciertos servidores luego de visitar las páginas alojadas en estos.
   d) Definir la prioridad de rastreo de las URLs, determinando el orden en que los crawlers visitan y procesan cada página web.

18. El concepto de *rank sink* en el algoritmo de PageRank representa páginas web:

a) Con distribuciones uniformes de la puntuación de PageRank.

b) Que tienen una puntuación de PageRank más alta que otras debido a la manipulación de enlaces entrantes y salientes.

c) Con una baja calidad de contenido y una cantidad insuficiente de enlaces salientes, lo que las hace menos relevantes en los resultados de búsqueda.

d) Con un alto número de enlaces salientes que no reciben enlaces entrantes, lo que puede afectar negativamente su puntuación de PageRank.

19. En el contexto de la RI en redes sociales se puede afirmar que:

a) El análisis de redes sociales se centra exclusivamente en la cantidad de seguidores que tiene un usuario en particular para determinar su influencia en la red.

b) El análisis de centralidad de intermediación se utiliza para identificar usuarios que son importantes en una red social debido a su posición como "puentes" entre diferentes grupos de usuarios.

c) El análisis de sentimientos se utiliza para determinar la popularidad de una publicación en redes sociales sin tener en cuenta la opinión de los usuarios.

d) El análisis de redes sociales no es útil para comprender la difusión de información en una red social específica, para ello se utiliza la medida de centralidad de vector propio.

20. Un investigador necesita recopilar datos de múltiples sitios web para un estudio académico, pero se enfrenta a varios desafíos al realizar el proceso de Web Scraping de manera ética y legal. ¿Cuál de las siguientes opciones describe mejor uno de los desafíos asociados al proceso de Web Scraping?

a) La necesidad de comprender la estructura del sitio web y su código HTML para extraer los datos correctamente.

b) La disponibilidad limitada de datos en línea que se pueden extraer utilizando técnicas de Web Scraping.

c) La dificultad para encontrar herramientas de Web Scraping gratuitas y fiables.

d) La necesidad de estar montado sobre un scrawler que cumpla con todas las políticas.

Trabajo de Control
Sistemas de Recuperación de Información
Fecha: 15 de julio de 2024
Temario #32

Nombre: _____ Grupo: _____

Seleccione en cada caso la(s) respuesta(s) correcta(s). Considere también no marcar ninguna opción.

1. En el algoritmo de Indexación Basada en Clasificación Bloqueada (BSBI), ¿cuál es el paso final para crear un índice invertido para la colección completa de los datos?
   a) Eliminar los términos duplicados de los índices.
   b) Indexar cada bloque de forma independiente.
   c) Fusionar los índices invertidos de cada bloque.
   d) Dividir la colección de datos en bloques de tamaño fijo.

2. Un sistema cuenta con la siguiente información:
   - Juan pertenece a la marina.
   - Juan es capellán (sacerdote encargado de una tarea específica fuera de la parroquia).
   - Los infantes de la marina suelen ser bebedores de cerveza.
   - Un capellán no suele ser bebedor de cerveza.
   - Un bebedor de cerveza suele tener sobrepeso.
   - Por lo general, un infante de la marina está en buena condición física.

   Si se representa la información en una red de herencia se puede concluir que:
   a) No se puede asegurar que Juan sea capellán.
   b) No se puede asegurar que Juan esté en buena condición física.
   c) Juan tiene sobrepeso.
   d) Existen dos razonamientos cancelables.

3. Se puede afirmar que:
   a) El término "Big Data" se refiere exclusivamente al volumen de datos que una organización maneja, sin tener en cuenta la velocidad, la variedad y la veracidad de los datos.
   b) Los SRI pueden beneficiarse de MapReduce para mejorar la RI relevante.
   c) MapReduce es un modelo de procesamiento distribuido utilizado para trabajar con grandes volúmenes de datos.
   d) Uno de los desafíos en el procesamiento de Big Data es la capacidad de gestionar y analizar datos provenientes de diversas fuentes y en diferentes formatos de manera eficiente.

4. Implementar índices invertidos en un SRI asegura:
   a) Acelerar el proceso de búsqueda al permitir búsquedas directas por contenido en lugar de por título.
   b) Reducir la cantidad de espacio de almacenamiento necesario al comprimir los datos de los documentos.
   c) Facilitar la búsqueda sobre los datos que contienen términos específicos al mantener una lista de datos para cada término único.
   d) Incrementar la seguridad de los datos almacenados al dificultar el acceso directo a la información sin el índice correcto.

5. En un sistema donde el conocimiento está definido a partir de reglas se puede asegurar que:
   a) El nivel de especificidad de las reglas está limitado.
   b) La representación de conocimiento está basada en la lógica proposicional.
   c) El razonamiento se activa en cadena hacia delante.
   d) El orden en que se definen las reglas no altera el razonamiento del sistema.

6. ¿Qué estrategia utilizan los Web Crawlers para asegurar un rastreo eficiente y respetuoso de los recursos de los sitios web?
   a) Ignorar completamente el archivo `robots.txt` de los sitios web.
   b) Visitar y rastrear todos los enlaces de una página web simultáneamente.
   c) Extraer únicamente contenido multimedia para reducir la carga en los servidores web.
   d) Seguir las directrices del archivo `robots.txt` y aplicar un retraso entre las solicitudes.

7. La política de ordenación de URLs en los Web Crawlers tiene como aspecto fundamental:
   a) Limitar el acceso de los crawlers a ciertas secciones de un sitio web, evitando el rastreo de URLs consideradas menos importantes o sensibles.
   b) Definir la estructura de la URL de destino, asegurando que estén ordenadas alfabéticamente para facilitar la navegación y la indexación.
   c) Establecer la prioridad de rastreo de las URLs, determinando el orden en que los crawlers visitan y procesan cada página web.

d) Determinar la forma en que los crawlers asignan un valor de relevancia a cada URL para clasificarlas en el índice de búsqueda.

8. El componente responsable de la gestión de recursos y planificación de tareas en Hadoop es:
   a) MapReduce.
   b) Hadoop Common.
   c) HDFS.
   d) YARN.

9. Sobre el algoritmo de PageRank, visto en clase, se puede afirmar que:
   a) Evalúa la importancia de un sitio web en función de la calidad y cantidad de enlaces entrantes que recibe de otros sitios web.
   b) Solo tiene en cuenta el contenido en un sitio web para determinar su relevancia en los resultados de búsqueda.
   c) Asigna una puntuación alta a los sitios web que tienen un gran número de enlaces entrantes sin tener en cuenta la calidad de esos enlaces.
   d) Asigna una puntuación baja a los sitios web que contienen muchos enlaces salientes, ya que indica una falta de relevancia.

10. En una biblioteca digital se necesita implementar un sistema que permita a los usuarios encontrar libros y artículos científicos de forma rápida y precisa. Los documentos están en diversos formatos, incluyendo PDF, EPUB y HTML. Se requiere seleccionar un algoritmo de indexación adecuado para el sistema, por lo que el programador designado para la implementación debe considerar:
    a) La velocidad de indexación y recuperación de datos.
    b) La capacidad para manejar documentos en diferentes formatos.
    c) La complejidad del algoritmo en términos de implementación y mantenimiento.
    d) La capacidad del algoritmo para procesar imágenes incrustadas o referenciadas en los ficheros.

11. Se tiene una red de ingredientes donde cada uno representa un nodo y las aristas simbolizan que los ingredientes forman parte de una misma receta. Se busca mejorar la experiencia culinaria mediante la elaboración de combinaciones de ingredientes más interesantes y creativas, para lo cual se debe:
    a) No considerar la centralidad de grado de los ingredientes, ya que todas las combinaciones de ingredientes son igualmente válidas.
    b) Utilizar la centralidad de grado para identificar los ingredientes menos conectados en la red y tomarlos en cuenta para su inclusión en las combinaciones.
    c) No tener en cuenta la centralidad de grado de los ingredientes para no darle mayor importancia a los ingredientes más comunes.
    d) Utilizar la centralidad de grado para identificar los ingredientes más populares en la red y crear combinaciones que incluyan una variedad de ingredientes menos comunes.

12. ¿Cuál de las siguientes opciones describe mejor la diferencia clave entre Web Crawling y Web Scraping?
    a) Web Crawling se centra en la exploración y recopilación de enlaces de múltiples sitios web, mientras que Web Scraping se enfoca en la extracción específica de datos de páginas web individuales.
    b) Web Crawling se realiza utilizando herramientas de automatización como Selenium WebDriver, mientras que Web Scraping se lleva a cabo mediante el análisis de HTML y CSS.
    c) Web Scraping implica el análisis de la estructura y el contenido de las páginas web para extraer datos, mientras que Web Crawling se refiere a la descarga y almacenamiento de páginas web completas.
    d) Web Scraping es más eficaz para rastrear e indexar contenido web para motores de búsqueda, mientras que Web Crawling se utiliza principalmente para la extracción de datos en proyectos de investigación.

13. Al "relajar" el concepto de clique en la detección de comunidades se intenta solucionar:
    a) La imposibilidad de conectar nodos distantes.
    b) La necesidad de datos externos para analizar la red.
    c) La uniformidad de los nodos en términos de grado.
    d) El solapamiento y la complejidad computacional.

14. ¿Qué es un "Uniform Resource Locator (URL) Frontier" en el contexto de Web Crawling?
    a) Un protocolo que define cómo se deben formatear las URLs para el crawling.
    b) Una lista prioritaria de URLs que aún no han sido visitadas por el crawler.
    c) Una técnica para filtrar URLs irrelevantes y mejorar la eficiencia del crawling.
    d) Una base de datos que almacena URLs únicas identificadas como recursos en la web.

15. En un sistema de control de tráfico urbano basado en reglas, ¿cuál de las siguientes reglas sería más efectiva para manejar situaciones de congestión en una intersección?
    a) Si hay pocos vehículos en la intersección, reducir el tiempo de los semáforos en verde.
    b) Si hay muchos vehículos en la intersección, aumentar el tiempo de los semáforos en verde.
    c) Si hay un vehículo de emergencia en la intersección, detener todos los demás vehículos.

d) Si hay muchos peatones cruzando la intersección, reducir el tiempo de los semáforos en rojo.

16. La Web 1.0 se caracteriza por:
    a) Las actualizaciones del contenido de los sitios web se efectúa de forma manual dentro del código HTML
    b) Los sitios web se centran en brindar información en lugar de facilitar la colaboración o participación de los usuarios.
    c) Uso de comunidades virtuales para popularizar los sitios web de noticias.
    d) Los propietarios de los sitios web proporcionan contenido de forma periódica.
17. En el contexto de la RI en redes sociales se puede afirmar que:
    a) El análisis de redes sociales se centra exclusivamente en la cantidad de seguidores que tiene un usuario en particular para determinar su influencia en la red.
    b) El análisis de centralidad de intermediación se utiliza para identificar usuarios que son importantes en una red social debido a su posición como "puentes" entre diferentes grupos de usuarios.
    c) El análisis de sentimientos se utiliza para determinar la popularidad de una publicación en redes sociales sin tener en cuenta la opinión de los usuarios.
    d) El análisis de redes sociales no es útil para comprender la difusión de información en una red social específica, para ello se utiliza la medida de centralidad de vector propio.
18. El posicionamiento de un sitio web en los motores de búsqueda puede ser afectado por:
    a) Utilizar técnicas de encubrimiento para mostrar contenido diferente a lo indexado por los motores de búsqueda.
    b) Contenido sin valor en el sitio web.
    c) Mantener una estructura de URL clara y coherente.
    d) Obtener enlaces de sitios web irrelevantes y de baja calidad.
19. En una red de transporte donde cada nodo es una parada de autobús y las aristas representan si existe un carro que pasa por ambos sitios, ¿qué puede mejorar el sistema de transporte?
    a) No considerar la centralidad de intermediación de las estaciones de transporte público al planificar rutas y horarios, ya que no tiene impacto en las conexiones entre las paradas.
    b) Utilizar el grafo inducido de los nodos con mayor valor en la centralidad de grado para aplicar la centralidad de intermediación con el propósito de reforzar las paradas con mayor tráfico.
    c) Utilizar la centralidad de intermediación para identificar las paradas de transferencia clave y establecer nuevas conexiones entre diferentes líneas de transporte público.
    d) No implementar sistemas de información en tiempo real para los usuarios, ya que pueden aumentar la carga de los trabajadores.
20. Dentro del ecosistema de Hadoop, el HDFS se caracteriza por:
    a) El modelo de acceso y de escritura de datos en tiempo real.
    b) La tolerancia a fallos mediante la replicación de datos.
    c) El almacenamiento exclusivo para archivos de texto.
    d) La capacidad ilimitada de almacenamiento.

Trabajo de Control
Sistemas de Recuperación de Información
Fecha: 15 de julio de 2024
Temario #33

Nombre: _____ Grupo: _____

Seleccione en cada caso la(s) respuesta(s) correcta(s). Considere también no marcar ninguna opción.

1. En un sistema de control de tráfico urbano basado en reglas, ¿cuál de las siguientes reglas sería más efectiva para manejar situaciones de congestión en una intersección?
   a) Si hay pocos vehículos en la intersección, reducir el tiempo de los semáforos en verde.
   b) Si hay muchos vehículos en la intersección, aumentar el tiempo de los semáforos en verde.
   c) Si hay un vehículo de emergencia en la intersección, detener todos los demás vehículos.
   d) Si hay muchos peatones cruzando la intersección, reducir el tiempo de los semáforos en rojo.

2. Con respecto a WordNet y su contribución a las ontologías y la representación del conocimiento en los sistemas de RI, puede afirmarse que:
   a) En WordNet, una palabra está asociada a un conjunto de sinónimos (synsets), siendo estas palabras intercambiables en un contexto.
   b) La integración de WordNet en sistemas de RI permite la expansión de consultas y la mejora de la precisión de los resultados al entender mejor el significado de los términos de búsqueda a través de su contexto semántico y las relaciones entre palabras.
   c) WordNet diferencia claramente entre los significados de palabras según su uso en diferentes contextos, lo que permite aplicaciones avanzadas en desambiguación semántica más allá de los sistemas de recomendación y RI.
   d) Una limitación de WordNet en la representación del conocimiento es su enfoque en el idioma inglés, lo que plantea desafíos en la aplicación global y la interoperabilidad con sistemas de información multilingües.

3. Una plataforma de comercio electrónico desea mejorar la experiencia del usuario al permitir una navegación más personalizada y contextualizada. Actualmente los usuarios tienen dificultades para encontrar productos relevantes debido a la gran cantidad de opciones disponibles. La empresa está interesada en implementar características de la Web 2.5 y la Web Semántica para abordar este problema. ¿Qué características podrían ayudar para ofrecer una navegación más personalizada y contextualizada?
   a) La implementación de ontologías y metadatos para enriquecer la descripción de productos y mejorar la precisión de las recomendaciones.
   b) La optimización de la velocidad de carga del sitio web para mejorar la experiencia del usuario y reducir el abandono del carrito de compra.
   c) La integración de redes sociales para permitir la recomendación de productos basada en las preferencias de amigos y contactos.
   d) La incorporación de los productos en tendencia en el mercado.

4. En el contexto del procesamiento de grandes conjuntos de datos, se puede asegurar que:
   a) MapReduce es un enfoque para procesar datos de forma secuencial en un solo servidor para evitar problemas de concurrencia.
   b) MapReduce divide una tarea en múltiples pasos de map y reduce que se ejecutan de forma secuencial en diferentes servidores para mejorar el rendimiento y la escalabilidad.
   c) MapReduce no es adecuado para el procesamiento de datos no estructurados.
   d) MapReduce solo puede manejar pequeñas cantidades de datos y no escala bien a grandes conjuntos de datos.

5. La integración de Hadoop y MapReduce en la RI trae como ventaja:
   a) La eliminación de la necesidad de sistemas de bases de datos.
   b) La posibilidad de la extracción de información relevante y la generación de resultados significativos de grandes conjuntos de datos.
   c) La garantía de la privacidad absoluta de los datos procesados.
   d) La reducción de los costos operativos a cero.

6. Al realizar la optimización de contenido para SEO debe considerarse:
   a) Seleccionar las palabras clave al azar.
   b) Utilizar etiquetas de título y meta descripciones únicas y relevantes para cada página.
   c) Incluir palabras clave de manera excesiva en el contenido para mejorar el posicionamiento.

d) Crear contenido valioso y original que satisfaga las necesidades de los usuarios.

7. La web actual se enfrenta a problemas como:
   a) Presencia de grandes volúmenes de datos estructurados.
   b) Presencia de una alta calidad en los datos.
   c) Presencia de datos volátiles y distribuidos.
   d) Heterogeneidad en los datos.

8. ¿Qué estrategia utilizan los Web Crawlers para asegurar un rastreo eficiente y respetuoso de los recursos de los sitios web?
   a) Ignorar completamente el archivo `robots.txt` de los sitios web.
   b) Visitar y rastrear todos los enlaces de una página web simultáneamente.
   c) Extraer únicamente contenido multimedia para reducir la carga en los servidores web.
   d) Seguir las directrices del archivo `robots.txt` y aplicar un retraso entre las solicitudes.

9. Considerando las dimensiones y desafíos inherentes a Big Data puede afirmarse, tomando en cuenta las características clave y las implicaciones para su procesamiento y análisis, que:
   a) Big Data se caracteriza principalmente por su pequeño volumen y uniformidad, permitiendo un procesamiento eficiente con mínimas adaptaciones de las herramientas de análisis de datos tradicionales.
   b) Big Data no desafía la capacidad de las herramientas tradicionales de procesamiento de datos para capturar, almacenar, gestionar y analizar efectivamente la información, dada la evolución constante de las capacidades computacionales y algoritmos de optimización.
   c) Aunque Big Data puede incluir datos estructurados, su naturaleza se expande al incorporar grandes cantidades de datos no estructurados y semiestructurados, lo que exige el uso de tecnologías especializadas en almacenamiento y procesamiento como Hadoop y sistemas de bases de datos NoSQL.
   d) Además de su complejidad y diversidad, Big Data introduce desafíos significativos en términos de veracidad y variabilidad al requerir métodos avanzados de limpieza y validación de datos para asegurar la integridad del análisis.

10. Una buena práctica de SEO para mejorar el posicionamiento de un sitio web en los motores de búsqueda es:
    a) Obtener enlaces de otros sitios web relevantes y de calidad que apunten al sitio.
    b) Copiar contenido directamente de otros sitios web populares para aumentar la cantidad de páginas indexadas.
    c) Utilizar técnicas de encubrimiento para mostrar contenido diferente a lo indexado por los motores de búsqueda.
    d) Llenar el contenido con palabras clave irrelevantes para aumentar la densidad de palabras clave.

11. El propósito de la política de revisitado en los Web Crawlers es:
    a) Determinar la frecuencia con la que los crawlers deben volver a visitar una URL específica para mantener la información actualizada en el índice de búsqueda.
    b) Establecer reglas sobre el tiempo máximo que los crawlers pueden pasar en un sitio web durante cada visita para evitar sobrecargar los servidores.
    c) Limitar el acceso de los crawlers a ciertos servidores luego de visitar las páginas alojadas en estos.
    d) Definir la prioridad de rastreo de las URLs, determinando el orden en que los crawlers visitan y procesan cada página web.

12. Analizar una red permite:
    a) Detectar posibles tendencias antes de que se conviertan en tendencia.
    b) Obtener predicciones exactas de eventos futuros en mercados financieros.
    c) Evaluar la calidad del contenido de un sitio web.
    d) Encontrar nodos “sensibles” o críticos para la red.

13. Se tiene un grafo donde cada nodo es un personaje de cierto libro de cuentos y la existencia de las aristas está definida si dos personajes aparecen en el mismo cuento. Se puede asegurar que:
    a) La centralidad de intermediación identifica a los personajes que actúan como conectores entre personajes de cuentos distintos.
    b) La centralidad de grado es útil para identificar los personajes con más conexiones dentro de la red, lo que puede indicar su importancia en el libro.
    c) La centralidad de cercanía ofrece una relación entre la cantidad de vecinos de un nodo con respecto a la longitud máxima de un camino dentro del grafo partiendo del nodo en cuestión.
    d) La centralidad de cercanía indica el grado de conexión de cada personaje con el resto de los personajes de los cuentos del libro.

14. Considerando las prácticas éticas y legales en el Web Scraping, se puede asegurar que:
    a) Web Scraping requiere considerar las políticas de `robots.txt` del sitio web objetivo.

b) Es importante revisar y respetar los términos de servicio del sitio web, así como las leyes aplicables de protección de datos y derechos de autor, antes de realizar Web Scraping.

c) Web Scraping sobre datos personales sin consentimiento es generalmente aceptado si los datos se utilizan con fines de investigación.

d) La extracción de datos mediante Web Scraping siempre es legal, independientemente de las leyes locales sobre derechos de autor y privacidad de datos definidos en los sitios web.

15. Sobre WordNet y su aplicación en el procesamiento del lenguaje natural, se puede afirmar que:

a) El diseño de WordNet facilita su integración en aplicaciones multilingües de NLP, aunque su desarrollo original se centró en el inglés.

b) Aunque WordNet es una herramienta valiosa en el NLP, su estructura no incluye información sobre la frecuencia de uso de las palabras en el lenguaje natural.

c) Los synsets facilitan la identificación de relaciones semánticas entre palabras, como la hiperonimia y la meronimia, enriqueciendo tareas de NLP.

d) WordNet proporciona una base para la desambiguación semántica al agrupar palabras con significados similares en synsets.

16. La diferencia entre la indexación por tokens y la indexación por conceptos puede definirse como:

a) La indexación por tokens asigna pesos a los términos basados en su importancia relativa, mientras que la indexación por conceptos utiliza un sistema de etiquetado para asociar términos con características generales.

b) La indexación por tokens divide los datos en términos individuales, mientras que la indexación por conceptos agrupa los datos en categorías definidas.

c) La indexación por tokens asigna un valor numérico a cada término de los datos, mientras que la indexación por conceptos utiliza algoritmos de encriptación para proteger la privacidad de los datos.

d) La indexación por tokens normaliza los datos reduciéndolos a su forma básica, mientras que la indexación por conceptos utiliza un método de ordenación para organizar los términos característicos de los datos.

17. Se tiene una red de ingredientes donde cada uno representa un nodo y las aristas simbolizan que los ingredientes forman parte de una misma receta. Se busca mejorar la experiencia culinaria mediante la elaboración de combinaciones de ingredientes más interesantes y creativas, para lo cual se debe:

a) No considerar la centralidad de grado de los ingredientes, ya que todas las combinaciones de ingredientes son igualmente válidas.

b) Utilizar la centralidad de grado para identificar los ingredientes menos conectados en la red y tomarlos en cuenta para su inclusión en las combinaciones.

c) No tener en cuenta la centralidad de grado de los ingredientes para no darle mayor importancia a los ingredientes más comunes.

d) Utilizar la centralidad de grado para identificar los ingredientes más populares en la red y crear combinaciones que incluyan una variedad de ingredientes menos comunes.

18. En la RI en el contexto de Big Data, se puede asegurar que:

a) El almacenamiento distribuido centraliza todos los datos en un único servidor para facilitar su gestión y mantenimiento.

b) El almacenamiento distribuido ofrece ventajas significativas en términos de escalabilidad y rendimiento en comparación con el almacenamiento centralizado.

c) El almacenamiento distribuido reparte los datos en múltiples servidores para mejorar la disponibilidad y la redundancia del sistema.

d) El almacenamiento distribuido es una técnica obsoleta.

19. Si se tiene el conjunto de páginas interconectadas $\{A \to B, C; B \to C; C \to A\}$, entonces la página con valor más alto de PageRank es:

a) A y B.

b) A.

c) C.

d) B.

20. Sobre el algoritmo de PageRank, visto en clase, se puede afirmar que:

a) Evalúa la importancia de un sitio web en función de la calidad y cantidad de enlaces entrantes que recibe de otros sitios web.

b) Solo tiene en cuenta el contenido en un sitio web para determinar su relevancia en los resultados de búsqueda.

c) Asigna una puntuación alta a los sitios web que tienen un gran número de enlaces entrantes sin tener en cuenta la calidad de esos enlaces.

d) Asigna una puntuación baja a los sitios web que contienen muchos enlaces salientes, ya que indica una falta de relevancia.

Trabajo de Control
Sistemas de Recuperación de Información
Fecha: 15 de julio de 2024
Temario #34

Nombre: _____ Grupo: _____

Seleccione en cada caso la(s) respuesta(s) correcta(s). Considere también no marcar ninguna opción.

1. Sobre el algoritmo de PageRank, visto en clase, se puede afirmar que:
   a) Evalúa la importancia de un sitio web en función de la calidad y cantidad de enlaces entrantes que recibe de otros sitios web.
   b) Solo tiene en cuenta el contenido en un sitio web para determinar su relevancia en los resultados de búsqueda.
   c) Asigna una puntuación alta a los sitios web que tienen un gran número de enlaces entrantes sin tener en cuenta la calidad de esos enlaces.
   d) Asigna una puntuación baja a los sitios web que contienen muchos enlaces salientes, ya que indica una falta de relevancia.

2. Se puede afirmar que:
   a) No es posible identificar subgrupos dentro de una red utilizando análisis de redes.
   b) La cantidad de conexiones de un nodo siempre indica su influencia en la red.
   c) El tamaño de una red es siempre indicativo de su efectividad en la transmisión de información.
   d) Todas las relaciones en una red tienen la misma importancia para el análisis.

3. Una plataforma de comercio electrónico desea mejorar la experiencia del usuario al permitir una navegación más personalizada y contextualizada. Actualmente los usuarios tienen dificultades para encontrar productos relevantes debido a la gran cantidad de opciones disponibles. La empresa está interesada en implementar características de la Web 2.5 y la Web Semántica para abordar este problema. ¿Qué características podrían ayudar para ofrecer una navegación más personalizada y contextualizada?
   a) La implementación de ontologías y metadatos para enriquecer la descripción de productos y mejorar la precisión de las recomendaciones.
   b) La optimización de la velocidad de carga del sitio web para mejorar la experiencia del usuario y reducir el abandono del carrito de compra.
   c) La integración de redes sociales para permitir la recomendación de productos basada en las preferencias de amigos y contactos.
   d) La incorporación de los productos en tendencia en el mercado.

4. ¿Qué algoritmos permiten obtener información de una red?
   a) Índices de centralidad.
   b) Detección de comunidades.
   c) Hypertext Induced Topic Selection (HITS).
   d) PageRank.

5. La web actual se enfrenta a problemas como:
   a) Presencia de grandes volúmenes de datos estructurados.
   b) Presencia de una alta calidad en los datos.
   c) Presencia de datos volátiles y distribuidos.
   d) Heterogeneidad en los datos.

6. En un sistema donde el conocimiento está orientado a objetos se puede asegurar que:
   a) Se enfatiza la atención a la información de la cual se extrajo el conocimiento.
   b) Los marcos y las bandas son las estructuras utilizadas para la representación del modelo.
   c) No es posible definir especializaciones de los objetos de la vida real dentro del sistema.
   d) Las funciones de agregación dificultan poder establecer relaciones entre los objetos.

7. En una red de transporte donde cada nodo es una parada de autobús y las aristas representan si existe un carro que pasa por ambos sitios, ¿qué puede mejorar el sistema de transporte?
   a) No considerar la centralidad de intermediación de las estaciones de transporte público al planificar rutas y horarios, ya que no tiene impacto en las conexiones entre las paradas.
   b) Utilizar el grafo inducido de los nodos con mayor valor en la centralidad de grado para aplicar la centralidad de intermediación con el propósito de reforzar las paradas con mayor tráfico.
   c) Utilizar la centralidad de intermediación para identificar las paradas de transferencia clave y establecer nuevas conexiones entre diferentes líneas de transporte público.

d) No implementar sistemas de información en tiempo real para los usuarios, ya que pueden aumentar la carga de los trabajadores.

8. Al realizar la optimización de contenido para SEO debe considerarse:
   a) Seleccionar las palabras clave al azar.
   b) Utilizar etiquetas de título y meta descripciones únicas y relevantes para cada página.
   c) Incluir palabras clave de manera excesiva en el contenido para mejorar el posicionamiento.
   d) Crear contenido valioso y original que satisfaga las necesidades de los usuarios.
9. La Web 3 se conoce como:
   a) Internet de las cosas.
   b) Web de solo lectura.
   c) Web semántica.
   d) Web de escritura-lectura.
10. En un grafo una comunidad es:
   a) Un conjunto de nodos que no comparten ninguna similitud estructural o funcional entre sí.
   b) Un conjunto de nodos aislados.
   c) Un conjunto de nodos altamente conectados que forman un subgrafo completamente independiente del resto de la red.
   d) Un conjunto de nodos que están más densamente interconectados entre sí que con los nodos fuera del conjunto.
11. La computación evolutiva:
   a) No es aplicable en la RI debido a la complejidad de los algoritmos evolutivos.
   b) Solo puede manejar conjuntos de datos pequeños y no es escalable a grandes volúmenes de datos.
   c) Utiliza algoritmos para buscar soluciones óptimas en grandes espacios de búsqueda, lo que la hace adecuada para problemas de optimización en la RI.
   d) Es útil solo para problemas de clasificación de documentos y no para otras tareas de RI en general.
12. En el contexto de la RI en redes sociales se puede afirmar que:
   a) El análisis de redes sociales se centra exclusivamente en la cantidad de seguidores que tiene un usuario en particular para determinar su influencia en la red.
   b) El análisis de centralidad de intermediación se utiliza para identificar usuarios que son importantes en una red social debido a su posición como "puentes" entre diferentes grupos de usuarios.
   c) El análisis de sentimientos se utiliza para determinar la popularidad de una publicación en redes sociales sin tener en cuenta la opinión de los usuarios.
   d) El análisis de redes sociales no es útil para comprender la difusión de información en una red social específica, para ello se utiliza la medida de centralidad de vector propio.
13. Analizar una red permite:
   a) Detectar posibles tendencias antes de que se conviertan en tendencia.
   b) Obtener predicciones exactas de eventos futuros en mercados financieros.
   c) Evaluar la calidad del contenido de un sitio web.
   d) Encontrar nodos "sensibles" o críticos para la red.
14. La Web 2.0 se caracteriza por:
   a) La existencia de sitios web dinámicos e interactivos que permiten a los usuarios participar, comentar e interactuar tanto con los creadores de contenido como con otros usuarios.
   b) El uso de distintas tecnologías para crear experiencias web más interactivas y con mayor capacidad de respuesta.
   c) Las actualizaciones del contenido de los sitios web se efectúa de forma manual dentro del código HTML
   d) Las plataformas ofrecen experiencias a la medida, permitiendo a los usuarios personalizar sus perfiles, recibir recomendaciones ajustadas al contenido y participar en filtrado colaborativo.
15. En el contexto de la representación del conocimiento basada en herencia, ¿qué caracteriza a la herencia cancelable?
   a) La herencia es el resultado del razonamiento no transitivo.
   b) Las conclusiones no están determinadas y dependen del nodo de interés.
   c) Las propiedades heredadas siempre se mantienen y no pueden anularse.
   d) No existe ambigüedad en las conclusiones obtenidas.
16. El algoritmo Hypertext Induced Topic Selection (HITS) intenta buscar nodos especiales. Estos son conocidos como:
   a) Autoridades.
   b) Hubs.
   c) Centrales.
   d) Sensibles.
17. Implementar índices invertidos en un SRI asegura:
   a) Acelerar el proceso de búsqueda al permitir búsquedas directas por contenido en lugar de por título.
   b) Reducir la cantidad de espacio de almacenamiento necesario al comprimir los datos de los documentos.

c) Facilitar la búsqueda sobre los datos que contienen términos específicos al mantener una lista de datos para cada término único.

d) Incrementar la seguridad de los datos almacenados al dificultar el acceso directo a la información sin el índice correcto.

18. En el contexto del procesamiento de grandes conjuntos de datos, se puede asegurar que:

a) MapReduce es un enfoque para procesar datos de forma secuencial en un solo servidor para evitar problemas de concurrencia.

b) MapReduce divide una tarea en múltiples pasos de map y reduce que se ejecutan de forma secuencial en diferentes servidores para mejorar el rendimiento y la escalabilidad.

c) MapReduce no es adecuado para el procesamiento de datos no estructurados.

d) MapReduce solo puede manejar pequeñas cantidades de datos y no escala bien a grandes conjuntos de datos.

19. La afirmación que mejor describe la política de amabilidad en los Web Crawlers es:

a) Los crawlers se diseñan para acceder a sitios web sin restricciones y extraer datos de manera agresiva para su indexación.

b) La política de amabilidad de los Web Crawlers dicta que los crawlers deben priorizar ciertos tipos de contenido sobre otros, ignorando completamente ciertas páginas web.

c) La política de amabilidad establece pautas y reglas sobre cómo los crawlers deben interactuar con los sitios web para minimizar la carga del servidor y respetar las directivas de los administradores del sitio.

d) Los Web Crawlers son libres de recopilar datos de cualquier sitio web sin restricciones, independientemente de la cantidad de tráfico que generen.

20. Un sistema cuenta con la siguiente información:

- Juan pertenece a la marina.
- Juan es capellán (sacerdote encargado de una tarea específica fuera de la parroquia).
- Los infantes de la marina suelen ser bebedores de cerveza.
- Un capellán no suele ser bebedor de cerveza.
- Un bebedor de cerveza suele tener sobrepeso.
- Por lo general, un infante de la marina está en buena condición física.

Si se representa la información en una red de herencia se puede concluir que:

a) No se puede asegurar que Juan sea capellán.

b) No se puede asegurar que Juan esté en buena condición física.

c) Juan tiene sobrepeso.

d) Existen dos razonamientos cancelables.

Trabajo de Control
Sistemas de Recuperación de Información
Fecha: 15 de julio de 2024
Temario #35

Nombre: _____ Grupo: _____

Seleccione en cada caso la(s) respuesta(s) correcta(s). Considere también no marcar ninguna opción.

1. Un SRI es capaz de:
   a) Crear los índices asociados a los datos sin tener que analizar cada dato.
   b) Generar índices invertidos de manera óptima sin considerar el contexto.
   c) Reducir el tiempo de indexación de los datos si utiliza servidores distribuidos dentro de la red para que cada uno ejecute la indexación del mismo conjunto de datos.
   d) No necesitar de ningún almacenamiento externo para alojar los índices de los datos.

2. La integración de Hadoop y MapReduce en la RI trae como ventaja:
   a) La eliminación de la necesidad de sistemas de bases de datos.
   b) La posibilidad de la extracción de información relevante y la generación de resultados significativos de grandes conjuntos de datos.
   c) La garantía de la privacidad absoluta de los datos procesados.
   d) La reducción de los costos operativos a cero.

3. Se puede afirmar que:
   a) El término "Big Data" se refiere exclusivamente al volumen de datos que una organización maneja, sin tener en cuenta la velocidad, la variedad y la veracidad de los datos.
   b) Los SRI pueden beneficiarse de MapReduce para mejorar la RI relevante.
   c) MapReduce es un modelo de procesamiento distribuido utilizado para trabajar con grandes volúmenes de datos.
   d) Uno de los desafíos en el procesamiento de Big Data es la capacidad de gestionar y analizar datos provenientes de diversas fuentes y en diferentes formatos de manera eficiente.

4. La afirmación que mejor refleja el principio subyacente de PageRank, considerando su importancia en la clasificación de los sitios web, es:
   a) PageRank valora más la cantidad de enlaces entrantes a una página web, independientemente de la calidad o relevancia de estos enlaces.
   b) La efectividad de PageRank se basa exclusivamente en el análisis de las palabras clave contenidas en los enlaces entrantes, sin considerar la estructura de enlace de la Web.
   c) El algoritmo de PageRank considera tanto la cantidad como la calidad de los enlaces entrantes, asignando mayor valor a los enlaces provenientes de sitios web considerados como "importantes".
   d) PageRank opera bajo el supuesto de que los enlaces entrantes y salientes tienen el mismo impacto en la valoración de la relevancia de una página web.

5. El posicionamiento de un sitio web en los motores de búsqueda puede ser afectado por:
   a) Utilizar técnicas de encubrimiento para mostrar contenido diferente a lo indexado por los motores de búsqueda.
   b) Contenido sin valor en el sitio web.
   c) Mantener una estructura de URL clara y coherente.
   d) Obtener enlaces de sitios web irrelevantes y de baja calidad.

6. ¿Qué estrategia utilizan los Web Crawlers para asegurar un rastreo eficiente y respetuoso de los recursos de los sitios web?
   a) Ignorar completamente el archivo `robots.txt` de los sitios web.
   b) Visitar y rastrear todos los enlaces de una página web simultáneamente.
   c) Extraer únicamente contenido multimedia para reducir la carga en los servidores web.
   d) Seguir las directrices del archivo `robots.txt` y aplicar un retraso entre las solicitudes.

7. Los Web Crawlers se enfrentan a desafíos constantes. Dentro de ellos se encuentran:
   a) La dificultad para generar el contenido dinámico en tiempo real.
   b) La modificación del código y la estructura del sitio web.
   c) La incapacidad para interpretar correctamente el lenguaje de programación utilizado en el desarrollo de los sitios web.
   d) La falta de acceso a la base de datos del servidor web para extraer información actualizada.

8. El propósito de la política de revisitado en los Web Crawlers es:
   a) Determinar la frecuencia con la que los crawlers deben volver a visitar una URL específica para mantener la información actualizada en el índice de búsqueda.
   b) Establecer reglas sobre el tiempo máximo que los crawlers pueden pasar en un sitio web durante cada visita para evitar sobrecargar los servidores.
   c) Limitar el acceso de los crawlers a ciertos servidores luego de visitar las páginas alojadas en estos.
   d) Definir la prioridad de rastreo de las URLs, determinando el orden en que los crawlers visitan y procesan cada página web.

9. En un sistema donde el conocimiento está definido a partir de reglas se puede asegurar que:
   a) El nivel de especificidad de las reglas está limitado.
   b) La representación de conocimiento está basada en la lógica proposicional.
   c) El razonamiento se activa en cadena hacia delante.
   d) El orden en que se definen las reglas no altera el razonamiento del sistema.

10. Al realizar la optimización de contenido para SEO debe considerarse:
    a) Seleccionar las palabras clave al azar.
    b) Utilizar etiquetas de título y meta descripciones únicas y relevantes para cada página.
    c) Incluir palabras clave de manera excesiva en el contenido para mejorar el posicionamiento.
    d) Crear contenido valioso y original que satisfaga las necesidades de los usuarios.

11. La diferencia entre la indexación por tokens y la indexación por conceptos puede definirse como:
    a) La indexación por tokens asigna pesos a los términos basados en su importancia relativa, mientras que la indexación por conceptos utiliza un sistema de etiquetado para asociar términos con características generales.
    b) La indexación por tokens divide los datos en términos individuales, mientras que la indexación por conceptos agrupa los datos en categorías definidas.
    c) La indexación por tokens asigna un valor numérico a cada término de los datos, mientras que la indexación por conceptos utiliza algoritmos de encriptación para proteger la privacidad de los datos.
    d) La indexación por tokens normaliza los datos reduciéndolos a su forma básica, mientras que la indexación por conceptos utiliza un método de ordenación para organizar los términos característicos de los datos.

12. Si se tiene el conjunto de páginas interconectadas $\{A \to B, C; B \to C; C \to A\}$, entonces la página con valor más alto de PageRank es:
    a) A y B.
    b) A.
    c) C.
    d) B.

13. El algoritmo Hypertext Induced Topic Selection (HITS) intenta buscar nodos especiales. Estos son conocidos como:
    a) Autoridades.
    b) Hubs.
    c) Centrales.
    d) Sensibles.

14. En una red de transporte donde cada nodo es una parada de autobús y las aristas representan si existe un carro que pasa por ambos sitios, ¿qué puede mejorar el sistema de transporte?
    a) No considerar la centralidad de intermediación de las estaciones de transporte público al planificar rutas y horarios, ya que no tiene impacto en las conexiones entre las paradas.
    b) Utilizar el grafo inducido de los nodos con mayor valor en la centralidad de grado para aplicar la centralidad de intermediación con el propósito de reforzar las paradas con mayor tráfico.
    c) Utilizar la centralidad de intermediación para identificar las paradas de transferencia clave y establecer nuevas conexiones entre diferentes líneas de transporte público.
    d) No implementar sistemas de información en tiempo real para los usuarios, ya que pueden aumentar la carga de los trabajadores.

15. Se tiene una red de ingredientes donde cada uno representa un nodo y las aristas simbolizan que los ingredientes forman parte de una misma receta. Se busca mejorar la experiencia culinaria mediante la elaboración de combinaciones de ingredientes más interesantes y creativas, para lo cual se debe:
    a) No considerar la centralidad de grado de los ingredientes, ya que todas las combinaciones de ingredientes son igualmente válidas.
    b) Utilizar la centralidad de grado para identificar los ingredientes menos conectados en la red y tomarlos en cuenta para su inclusión en las combinaciones.
    c) No tener en cuenta la centralidad de grado de los ingredientes para no darle mayor importancia a los ingredientes más comunes.
    d) Utilizar la centralidad de grado para identificar los ingredientes más populares en la red y crear combinaciones que incluyan una variedad de ingredientes menos comunes.

16. Implementar índices invertidos en un SRI asegura:
    a) Acelerar el proceso de búsqueda al permitir búsquedas directas por contenido en lugar de por título.
    b) Reducir la cantidad de espacio de almacenamiento necesario al comprimir los datos de los documentos.
    c) Facilitar la búsqueda sobre los datos que contienen términos específicos al mantener una lista de datos para cada término único.
    d) Incrementar la seguridad de los datos almacenados al dificultar el acceso directo a la información sin el índice correcto.

17. En la RI en el contexto de Big Data, se puede asegurar que:
    a) El almacenamiento distribuido centraliza todos los datos en un único servidor para facilitar su gestión y mantenimiento.
    b) El almacenamiento distribuido ofrece ventajas significativas en términos de escalabilidad y rendimiento en comparación con el almacenamiento centralizado.
    c) El almacenamiento distribuido reparte los datos en múltiples servidores para mejorar la disponibilidad y la redundancia del sistema.
    d) El almacenamiento distribuido es una técnica obsoleta.

18. La centralidad de intermediación de un nodo indica:
    a) La cantidad de nodos vecinos directos.
    b) La resistencia del nodo a fallos.
    c) El número total de conexiones entrantes y salientes.
    d) La frecuencia con la que un nodo actúa como puente en el camino más corto entre otros dos nodos.

19. La afirmación que mejor describe la política de amabilidad en los Web Crawlers es:
    a) Los crawlers se diseñan para acceder a sitios web sin restricciones y extraer datos de manera agresiva para su indexación.
    b) La política de amabilidad de los Web Crawlers dicta que los crawlers deben priorizar ciertos tipos de contenido sobre otros, ignorando completamente ciertas páginas web.
    c) La política de amabilidad establece pautas y reglas sobre cómo los crawlers deben interactuar con los sitios web para minimizar la carga del servidor y respetar las directivas de los administradores del sitio.
    d) Los Web Crawlers son libres de recopilar datos de cualquier sitio web sin restricciones, independientemente de la cantidad de tráfico que generen.

20. Los algoritmos para detectar comunidades en una red intentan:
    a) Buscar conjuntos donde cada nodo de un mismo conjunto tenga características similares al resto de los nodos del conjunto.
    b) Seleccionar aleatoriamente nodos de alto grado y sus vecinos.
    c) Buscar subgrafos tal que no incluyan nodos cuya ausencia desconecte al subgrafo.
    d) Encontrar grupos donde los nodos pertenecientes a los mismos grupos son cercanos bajo cierta métrica y lejanos con respecto a los nodos de otros grupos.

Trabajo de Control
Sistemas de Recuperación de Información
Fecha: 15 de julio de 2024
Temario #36

Nombre: _____ Grupo: _____

Seleccione en cada caso la(s) respuesta(s) correcta(s). Considere también no marcar ninguna opción.

1. La política de ordenación de URLs en los Web Crawlers tiene como aspecto fundamental:
   a) Limitar el acceso de los crawlers a ciertas secciones de un sitio web, evitando el rastreo de URLs consideradas menos importantes o sensibles.
   b) Definir la estructura de la URL de destino, asegurando que estén ordenadas alfabéticamente para facilitar la navegación y la indexación.
   c) Establecer la prioridad de rastreo de las URLs, determinando el orden en que los crawlers visitan y procesan cada página web.
   d) Determinar la forma en que los crawlers asignan un valor de relevancia a cada URL para clasificarlas en el índice de búsqueda.

2. En un grafo una comunidad es:
   a) Un conjunto de nodos que no comparten ninguna similitud estructural o funcional entre sí.
   b) Un conjunto de nodos aislados.
   c) Un conjunto de nodos altamente conectados que forman un subgrafo completamente independiente del resto de la red.
   d) Un conjunto de nodos que están más densamente interconectados entre sí que con los nodos fuera del conjunto.

3. La integración de Hadoop y MapReduce en la RI trae como ventaja:
   a) La eliminación de la necesidad de sistemas de bases de datos.
   b) La posibilidad de la extracción de información relevante y la generación de resultados significativos de grandes conjuntos de datos.
   c) La garantía de la privacidad absoluta de los datos procesados.
   d) La reducción de los costos operativos a cero.

4. La computación evolutiva:
   a) No es aplicable en la RI debido a la complejidad de los algoritmos evolutivos.
   b) Solo puede manejar conjuntos de datos pequeños y no es escalable a grandes volúmenes de datos.
   c) Utiliza algoritmos para buscar soluciones óptimas en grandes espacios de búsqueda, lo que la hace adecuada para problemas de optimización en la RI.
   d) Es útil solo para problemas de clasificación de documentos y no para otras tareas de RI en general.

5. La centralidad de intermediación de un nodo indica:
   a) La cantidad de nodos vecinos directos.
   b) La resistencia del nodo a fallos.
   c) El número total de conexiones entrantes y salientes.
   d) La frecuencia con la que un nodo actúa como puente en el camino más corto entre otros dos nodos.

6. La afirmación que mejor refleja el principio subyacente de PageRank, considerando su importancia en la clasificación de los sitios web, es:
   a) PageRank valora más la cantidad de enlaces entrantes a una página web, independientemente de la calidad o relevancia de estos enlaces.
   b) La efectividad de PageRank se basa exclusivamente en el análisis de las palabras clave contenidas en los enlaces entrantes, sin considerar la estructura de enlace de la Web.
   c) El algoritmo de PageRank considera tanto la cantidad como la calidad de los enlaces entrantes, asignando mayor valor a los enlaces provenientes de sitios web considerados como "importantes".
   d) PageRank opera bajo el supuesto de que los enlaces entrantes y salientes tienen el mismo impacto en la valoración de la relevancia de una página web.

7. ¿Qué estrategia utilizan los Web Crawlers para asegurar un rastreo eficiente y respetuoso de los recursos de los sitios web?
   a) Ignorar completamente el archivo `robots.txt` de los sitios web.
   b) Visitar y rastrear todos los enlaces de una página web simultáneamente.
   c) Extraer únicamente contenido multimedia para reducir la carga en los servidores web.
   d) Seguir las directrices del archivo `robots.txt` y aplicar un retraso entre las solicitudes.

8. Sobre WordNet y su aplicación en el procesamiento del lenguaje natural, se puede afirmar que:
   a) El diseño de WordNet facilita su integración en aplicaciones multilingües de NLP, aunque su desarrollo original se centró en el inglés.

b) Aunque WordNet es una herramienta valiosa en el NLP, su estructura no incluye información sobre la frecuencia de uso de las palabras en el lenguaje natural.

c) Los synsets facilitan la identificación de relaciones semánticas entre palabras, como la hiperonimia y la meronimia, enriqueciendo tareas de NLP.

d) WordNet proporciona una base para la desambiguación semántica al agrupar palabras con significados similares en synsets.

9. En un sistema donde el conocimiento está orientado a objetos se puede asegurar que:

a) Se enfatiza la atención a la información de la cual se extrajo el conocimiento.

b) Los marcos y las bandas son las estructuras utilizadas para la representación del modelo.

c) No es posible definir especializaciones de los objetos de la vida real dentro del sistema.

d) Las funciones de agregación dificultan poder establecer relaciones entre los objetos.

10. Un SRI es capaz de:

a) Crear los índices asociados a los datos sin tener que analizar cada dato.

b) Generar índices invertidos de manera óptima sin considerar el contexto.

c) Reducir el tiempo de indexación de los datos si utiliza servidores distribuidos dentro de la red para que cada uno ejecute la indexación del mismo conjunto de datos.

d) No necesitar de ningún almacenamiento externo para alojar los índices de los datos.

11. Para contribuir positivamente al posicionamiento orgánico de un sitio web en los motores de búsqueda se puede:

a) Crear contenido relevante y de alta calidad que satisfaga las necesidades de información de los usuarios.

b) Mejorar la velocidad de carga del sitio web y asegurar que sea *responsive* y fácil de usar en dispositivos móviles.

c) Incluir una densidad alta de palabras clave para asegurar que el sitio web aparezca en tantas búsquedas como sea posible.

d) Obtener enlaces entrantes de otros sitios web de alta autoridad y relevancia temática.

12. En un SRI la indexación:

a) Consiste en asociar un identificador único a cada dato almacenado en el sistema.

b) Mejora la experiencia del usuario.

c) Permite la organización y la categorización de la información.

d) Facilita la RI relevante.

13. Una plataforma de comercio electrónico desea mejorar la experiencia del usuario al permitir una navegación más personalizada y contextualizada. Actualmente los usuarios tienen dificultades para encontrar productos relevantes debido a la gran cantidad de opciones disponibles. La empresa está interesada en implementar características de la Web 2.5 y la Web Semántica para abordar este problema. ¿Qué características podrían ayudar para ofrecer una navegación más personalizada y contextualizada?

a) La implementación de ontologías y metadatos para enriquecer la descripción de productos y mejorar la precisión de las recomendaciones.

b) La optimización de la velocidad de carga del sitio web para mejorar la experiencia del usuario y reducir el abandono del carrito de compra.

c) La integración de redes sociales para permitir la recomendación de productos basada en las preferencias de amigos y contactos.

d) La incorporación de los productos en tendencia en el mercado.

14. El posicionamiento de un sitio web en los motores de búsqueda puede ser afectado por:

a) Utilizar técnicas de encubrimiento para mostrar contenido diferente a lo indexado por los motores de búsqueda.

b) Contenido sin valor en el sitio web.

c) Mantener una estructura de URL clara y coherente.

d) Obtener enlaces de sitios web irrelevantes y de baja calidad.

15. La afirmación que mejor describe la política de amabilidad en los Web Crawlers es:

a) Los crawlers se diseñan para acceder a sitios web sin restricciones y extraer datos de manera agresiva para su indexación.

b) La política de amabilidad de los Web Crawlers dicta que los crawlers deben priorizar ciertos tipos de contenido sobre otros, ignorando completamente ciertas páginas web.

c) La política de amabilidad establece pautas y reglas sobre cómo los crawlers deben interactuar con los sitios web para minimizar la carga del servidor y respetar las directivas de los administradores del sitio.

d) Los Web Crawlers son libres de recopilar datos de cualquier sitio web sin restricciones, independientemente de la cantidad de tráfico que generen.

16. Implementar índices invertidos en un SRI asegura:

a) Acelerar el proceso de búsqueda al permitir búsquedas directas por contenido en lugar de por título.

b) Reducir la cantidad de espacio de almacenamiento necesario al comprimir los datos de los documentos.

c) Facilitar la búsqueda sobre los datos que contienen términos específicos al mantener una lista de datos para cada término único.

d) Incrementar la seguridad de los datos almacenados al dificultar el acceso directo a la información sin el índice correcto.

17. Dentro del ecosistema de Hadoop, el HDFS se caracteriza por:

a) El modelo de acceso y de escritura de datos en tiempo real.

b) La tolerancia a fallos mediante la replicación de datos.

c) El almacenamiento exclusivo para archivos de texto.

d) La capacidad ilimitada de almacenamiento.

18. El algoritmo de PageRank puede describirse como un procedimiento utilizado para:

a) Determinar la velocidad de carga de una página web en un navegador.

b) Calcular la relevancia de una página web en función de la cantidad y calidad de los enlaces que apuntan hacia ella.

c) Prevenir el spam y el contenido no deseado en las páginas web.

d) Clasificar las páginas web en función de su edad y autoridad.

19. Existen varias estrategias consideradas como no recomendables para mejorar el SEO de un sitio web. Dentro de estas estrategias negativas, se encuentran:

a) Construir enlaces naturales de sitios web con autoridad y relevancia temática.

b) Evitar el uso de las meta etiquetas y descripciones del sitio web para reflejar el contenido de la página.

c) Incluir una cantidad excesiva de palabras clave irrelevantes para intentar manipular los rankings de búsqueda.

d) Utilizar técnicas de encubrimiento para mostrar contenido diferente a lo indexado por los motores de búsqueda.

20. Referente al Web Crawler puede afirmarse que:

a) Los hipervínculos encontrados en cada sitio web que no pertenecen al dominio donde fueron encontrados se desechan, puesto que no expande el conjunto de URLs sin visitar.

b) Las páginas visitadas no se procesan nunca más.

c) No necesita de un conjunto inicial de URLs para recorrer la Web.

d) No tiene como objetivo indexar y recopilar información de diferentes sitios web.

Trabajo de Control
Sistemas de Recuperación de Información
Fecha: 15 de julio de 2024
Temario #37

Nombre: _____ Grupo: _____

Seleccione en cada caso la(s) respuesta(s) correcta(s). Considere también no marcar ninguna opción.

1. En la RI en el contexto de Big Data, se puede asegurar que:
   a) El almacenamiento distribuido centraliza todos los datos en un único servidor para facilitar su gestión y mantenimiento.
   b) El almacenamiento distribuido ofrece ventajas significativas en términos de escalabilidad y rendimiento en comparación con el almacenamiento centralizado.
   c) El almacenamiento distribuido reparte los datos en múltiples servidores para mejorar la disponibilidad y la redundancia del sistema.
   d) El almacenamiento distribuido es una técnica obsoleta.

2. Considerando las prácticas éticas y legales en el Web Scraping, se puede asegurar que:
   a) Web Scraping requiere considerar las políticas de `robots.txt` del sitio web objetivo.
   b) Es importante revisar y respetar los términos de servicio del sitio web, así como las leyes aplicables de protección de datos y derechos de autor, antes de realizar Web Scraping.
   c) Web Scraping sobre datos personales sin consentimiento es generalmente aceptado si los datos se utilizan con fines de investigación.
   d) La extracción de datos mediante Web Scraping siempre es legal, independientemente de las leyes locales sobre derechos de autor y privacidad de datos definidos en los sitios web.

3. Un investigador necesita recopilar datos de múltiples sitios web para un estudio académico, pero se enfrenta a varios desafíos al realizar el proceso de Web Scraping de manera ética y legal. ¿Cuál de las siguientes opciones describe mejor uno de los desafíos asociados al proceso de Web Scraping?
   a) La necesidad de comprender la estructura del sitio web y su código HTML para extraer los datos correctamente.
   b) La disponibilidad limitada de datos en línea que se pueden extraer utilizando técnicas de Web Scraping.
   c) La dificultad para encontrar herramientas de Web Scraping gratuitas y fiables.
   d) La necesidad de estar montado sobre un scrawler que cumpla con todas las políticas.

4. En el contexto de la representación del conocimiento basada en herencia, ¿qué caracteriza a la herencia cancelable?
   a) La herencia es el resultado del razonamiento no transitivo.
   b) Las conclusiones no están determinadas y dependen del nodo de interés.
   c) Las propiedades heredadas siempre se mantienen y no pueden anularse.
   d) No existe ambigüedad en las conclusiones obtenidas.

5. Una plataforma de comercio electrónico desea mejorar la experiencia del usuario al permitir una navegación más personalizada y contextualizada. Actualmente los usuarios tienen dificultades para encontrar productos relevantes debido a la gran cantidad de opciones disponibles. La empresa está interesada en implementar características de la Web 2.5 y la Web Semántica para abordar este problema. ¿Qué características podrían ayudar para ofrecer una navegación más personalizada y contextualizada?
   a) La implementación de ontologías y metadatos para enriquecer la descripción de productos y mejorar la precisión de las recomendaciones.
   b) La optimización de la velocidad de carga del sitio web para mejorar la experiencia del usuario y reducir el abandono del carrito de compra.
   c) La integración de redes sociales para permitir la recomendación de productos basada en las preferencias de amigos y contactos.
   d) La incorporación de los productos en tendencia en el mercado.

6. Analizar una red permite:
   a) Detectar posibles tendencias antes de que se conviertan en tendencia.
   b) Obtener predicciones exactas de eventos futuros en mercados financieros.
   c) Evaluar la calidad del contenido de un sitio web.
   d) Encontrar nodos “sensibles” o críticos para la red.

7. Considerando las dimensiones y desafíos inherentes a Big Data puede afirmarse, tomando en cuenta las características clave y las implicaciones para su procesamiento y análisis, que:
   a) Big Data se caracteriza principalmente por su pequeño volumen y uniformidad, permitiendo un procesamiento eficiente con mínimas adaptaciones de las herramientas de análisis de datos tradicionales.
   b) Big Data no desafía la capacidad de las herramientas tradicionales de procesamiento de datos para capturar, almacenar, gestionar y analizar efectivamente la información, dada la evolución constante de las capacidades computacionales y algoritmos de optimización.
   c) Aunque Big Data puede incluir datos estructurados, su naturaleza se expande al incorporar grandes cantidades de datos no estructurados y semiestructurados, lo que exige el uso de tecnologías especializadas en almacenamiento y procesamiento como Hadoop y sistemas de bases de datos NoSQL.
   d) Además de su complejidad y diversidad, Big Data introduce desafíos significativos en términos de veracidad y variabilidad al requerir métodos avanzados de limpieza y validación de datos para asegurar la integridad del análisis.

8. Sobre WordNet y su aplicación en el procesamiento del lenguaje natural, se puede afirmar que:
   a) El diseño de WordNet facilita su integración en aplicaciones multilingües de NLP, aunque su desarrollo original se centró en el inglés.
   b) Aunque WordNet es una herramienta valiosa en el NLP, su estructura no incluye información sobre la frecuencia de uso de las palabras en el lenguaje natural.
   c) Los synsets facilitan la identificación de relaciones semánticas entre palabras, como la hiperonimia y la meronimia, enriqueciendo tareas de NLP.
   d) WordNet proporciona una base para la desambiguación semántica al agrupar palabras con significados similares en synsets.

9. Sobre el algoritmo de PageRank, visto en clase, se puede afirmar que:
   a) Evalúa la importancia de un sitio web en función de la calidad y cantidad de enlaces entrantes que recibe de otros sitios web.
   b) Solo tiene en cuenta el contenido en un sitio web para determinar su relevancia en los resultados de búsqueda.
   c) Asigna una puntuación alta a los sitios web que tienen un gran número de enlaces entrantes sin tener en cuenta la calidad de esos enlaces.
   d) Asigna una puntuación baja a los sitios web que contienen muchos enlaces salientes, ya que indica una falta de relevancia.

10. El conjunto frontera de URLs en un Web Crawler:
    a) Delimita los sitios web que visitará en futuras iteraciones del proceso.
    b) Es similar a un conjunto de URLs que esperan ser visitadas.
    c) Almacena hipervínculos que pertenecen al mismo dominio del conjunto semilla de URLs con que inició el crawler.
    d) Indica las secciones que no pueden ser visitadas de cada sitio web.

11. En el contexto del procesamiento de grandes conjuntos de datos, se puede asegurar que:
    a) MapReduce es un enfoque para procesar datos de forma secuencial en un solo servidor para evitar problemas de concurrencia.
    b) MapReduce divide una tarea en múltiples pasos de map y reduce que se ejecutan de forma secuencial en diferentes servidores para mejorar el rendimiento y la escalabilidad.
    c) MapReduce no es adecuado para el procesamiento de datos no estructurados.
    d) MapReduce solo puede manejar pequeñas cantidades de datos y no escala bien a grandes conjuntos de datos.

12. Al "relajar" el concepto de clique en la detección de comunidades se intenta solucionar:
    a) La imposibilidad de conectar nodos distantes.
    b) La necesidad de datos externos para analizar la red.
    c) La uniformidad de los nodos en términos de grado.
    d) El solapamiento y la complejidad computacional.

13. Si una red cumple la propiedad de ser un grafo de mundo pequeño, entonces se conoce que:
    a) El número de componentes fuertemente conexas está relacionado con la cantidad de grafos K_n presentes.
    b) La longitud media del camino entre todo par de vértices es pequeña.
    c) El grafo es un anillo regular de grado 5.
    d) La red posee pocos vértices.

14. ¿Cuál de las siguientes opciones describe mejor la diferencia clave entre Web Crawling y Web Scraping?
    a) Web Crawling se centra en la exploración y recopilación de enlaces de múltiples sitios web, mientras que Web Scraping se enfoca en la extracción específica de datos de páginas web individuales.

b) Web Crawling se realiza utilizando herramientas de automatización como Selenium WebDriver, mientras que Web Scraping se lleva a cabo mediante el análisis de HTML y CSS.

c) Web Scraping implica el análisis de la estructura y el contenido de las páginas web para extraer datos, mientras que Web Crawling se refiere a la descarga y almacenamiento de páginas web completas.

d) Web Scraping es más eficaz para rastrear e indexar contenido web para motores de búsqueda, mientras que Web Crawling se utiliza principalmente para la extracción de datos en proyectos de investigación.

15. La premisa básica del algoritmo de PageRank para clasificar páginas web en los resultados de búsqueda es:

a) La antigüedad de la página web es el principal factor para determinar su clasificación.

b) El contenido y la relevancia de las palabras clave en la página web determinan su posición.

c) Los enlaces entrantes a una página web desde otras páginas contribuyen a su importancia y clasificación.

d) La cantidad de visitas que recibe una página web determina su clasificación.

16. En un grafo una comunidad es:

a) Un conjunto de nodos que no comparten ninguna similitud estructural o funcional entre sí.

b) Un conjunto de nodos aislados.

c) Un conjunto de nodos altamente conectados que forman un subgrafo completamente independiente del resto de la red.

d) Un conjunto de nodos que están más densamente interconectados entre sí que con los nodos fuera del conjunto.

17. En la RI en el contexto de Big Data, se puede asegurar que:

a) La indexación distribuida divide los datos en múltiples fragmentos que se almacenan en varios nodos para permitir búsquedas paralelas y mejorar la escalabilidad.

b) La indexación distribuida no ofrece ventajas en términos de rendimiento y escalabilidad en comparación con la indexación centralizada.

c) La indexación distribuida es una técnica obsoleta.

d) La indexación distribuida almacena todos los datos en un solo servidor para facilitar su acceso y búsqueda.

18. ¿Qué estrategia utilizan los Web Crawlers para asegurar un rastreo eficiente y respetuoso de los recursos de los sitios web?

a) Ignorar completamente el archivo `robots.txt` de los sitios web.

b) Visitar y rastrear todos los enlaces de una página web simultáneamente.

c) Extraer únicamente contenido multimedia para reducir la carga en los servidores web.

d) Seguir las directrices del archivo `robots.txt` y aplicar un retraso entre las solicitudes.

19. El concepto de *rank sink* en el algoritmo de PageRank representa páginas web:

a) Con distribuciones uniformes de la puntuación de PageRank.

b) Que tienen una puntuación de PageRank más alta que otras debido a la manipulación de enlaces entrantes y salientes.

c) Con una baja calidad de contenido y una cantidad insuficiente de enlaces salientes, lo que las hace menos relevantes en los resultados de búsqueda.

d) Con un alto número de enlaces salientes que no reciben enlaces entrantes, lo que puede afectar negativamente su puntuación de PageRank.

20. El posicionamiento de un sitio web en los motores de búsqueda puede ser afectado por:

a) Utilizar técnicas de encubrimiento para mostrar contenido diferente a lo indexado por los motores de búsqueda.

b) Contenido sin valor en el sitio web.

c) Mantener una estructura de URL clara y coherente.

d) Obtener enlaces de sitios web irrelevantes y de baja calidad.

Trabajo de Control
Sistemas de Recuperación de Información
Fecha: 15 de julio de 2024
Temario #38

Nombre: _____ Grupo: _____

Seleccione en cada caso la(s) respuesta(s) correcta(s). Considere también no marcar ninguna opción.

1. ¿Qué estrategia utilizan los Web Crawlers para asegurar un rastreo eficiente y respetuoso de los recursos de los sitios web?
   a) Ignorar completamente el archivo `robots.txt` de los sitios web.
   b) Visitar y rastrear todos los enlaces de una página web simultáneamente.
   c) Extraer únicamente contenido multimedia para reducir la carga en los servidores web.
   d) Seguir las directrices del archivo `robots.txt` y aplicar un retraso entre las solicitudes.

2. Existen varias estrategias consideradas como no recomendables para mejorar el SEO de un sitio web. Dentro de estas estrategias negativas, se encuentran:
   a) Construir enlaces naturales de sitios web con autoridad y relevancia temática.
   b) Evitar el uso de las meta etiquetas y descripciones del sitio web para reflejar el contenido de la página.
   c) Incluir una cantidad excesiva de palabras clave irrelevantes para intentar manipular los rankings de búsqueda.
   d) Utilizar técnicas de encubrimiento para mostrar contenido diferente a lo indexado por los motores de búsqueda.

3. No se considera como técnica para detectar comunidades en una red:
   a) Analizar la mutualidad de los enlaces.
   b) Usar el agrupamiento jerárquico.
   c) Utilizar el algoritmo de K-Means.
   d) Encontrar cliques de vértices de grado par.

4. Si se tiene el conjunto de páginas interconectadas $\{A \to B, C; B \to C; C \to A\}$, entonces la página con valor más alto de PageRank es:
   a) A y B.
   b) A.
   c) C.
   d) B.

5. El uso de Hadoop y MapReduce en el contexto de la RI tiene como objetivo:
   a) Automatizar el mantenimiento de sistemas de bases de datos.
   b) Facilitar el análisis en tiempo real de datos de redes sociales.
   c) Mejorar la eficiencia energética en centros de datos.
   d) Procesar y analizar grandes conjuntos de datos para la RI.

6. En el algoritmo de Indexación Basada en Clasificación Bloqueada (BSBI), ¿cuál es el paso final para crear un índice invertido para la colección completa de los datos?
   a) Eliminar los términos duplicados de los índices.
   b) Indexar cada bloque de forma independiente.
   c) Fusionar los índices invertidos de cada bloque.
   d) Dividir la colección de datos en bloques de tamaño fijo.

7. En una red de transporte donde cada nodo es una parada de autobús y las aristas representan si existe un carro que pasa por ambos sitios, ¿qué puede mejorar el sistema de transporte?
   a) No considerar la centralidad de intermediación de las estaciones de transporte público al planificar rutas y horarios, ya que no tiene impacto en las conexiones entre las paradas.
   b) Utilizar el grafo inducido de los nodos con mayor valor en la centralidad de grado para aplicar la centralidad de intermediación con el propósito de reforzar las paradas con mayor tráfico.
   c) Utilizar la centralidad de intermediación para identificar las paradas de transferencia clave y establecer nuevas conexiones entre diferentes líneas de transporte público.
   d) No implementar sistemas de información en tiempo real para los usuarios, ya que pueden aumentar la carga de los trabajadores.

8. ¿Qué algoritmos permiten obtener información de una red?
   a) Índices de centralidad.
   b) Detección de comunidades.
   c) Hypertext Induced Topic Selection (HITS).
   d) PageRank.

9. En una biblioteca digital se necesita implementar un sistema que permita a los usuarios encontrar libros y artículos científicos de forma rápida y precisa. Los documentos están en diversos formatos, incluyendo PDF, EPUB y HTML. Se requiere seleccionar un algoritmo de indexación adecuado para el sistema, por lo que el programador designado para la implementación debe considerar:
   a) La velocidad de indexación y recuperación de datos.
   b) La capacidad para manejar documentos en diferentes formatos.
   c) La complejidad del algoritmo en términos de implementación y mantenimiento.
   d) La capacidad del algoritmo para procesar imágenes incrustadas o referenciadas en los ficheros.

10. Analizar una red permite:
   a) Detectar posibles tendencias antes de que se conviertan en tendencia.
   b) Obtener predicciones exactas de eventos futuros en mercados financieros.
   c) Evaluar la calidad del contenido de un sitio web.
   d) Encontrar nodos "sensibles" o críticos para la red.

11. El posicionamiento de un sitio web en los motores de búsqueda puede ser afectado por:
   a) Utilizar técnicas de encubrimiento para mostrar contenido diferente a lo indexado por los motores de búsqueda.
   b) Contenido sin valor en el sitio web.
   c) Mantener una estructura de URL clara y coherente.
   d) Obtener enlaces de sitios web irrelevantes y de baja calidad.

12. En el contexto del procesamiento de grandes conjuntos de datos, se puede asegurar que:
   a) MapReduce es un enfoque para procesar datos de forma secuencial en un solo servidor para evitar problemas de concurrencia.
   b) MapReduce divide una tarea en múltiples pasos de map y reduce que se ejecutan de forma secuencial en diferentes servidores para mejorar el rendimiento y la escalabilidad.
   c) MapReduce no es adecuado para el procesamiento de datos no estructurados.
   d) MapReduce solo puede manejar pequeñas cantidades de datos y no escala bien a grandes conjuntos de datos.

13. Si una red cumple la propiedad de ser un grafo de mundo pequeño, entonces se conoce que:
   a) El número de componentes fuertemente conexas está relacionado con la cantidad de grafos K_n presentes.
   b) La longitud media del camino entre todo par de vértices es pequeña.
   c) El grafo es un anillo regular de grado 5.
   d) La red posee pocos vértices.

14. Para transformar el contenido no estructurado de las páginas web en datos estructurados, el Web Scraping:
   a) Extrae información basada en patrones de HTML/CSS.
   b) Analiza los protocolos de red.
   c) Interpreta el código JavaScript en tiempo real.
   d) Convierte de forma automática imágenes a texto.

15. Con respecto a WordNet y su contribución a las ontologías y la representación del conocimiento en los sistemas de RI, puede afirmarse que:
   a) En WordNet, una palabra está asociada a un conjunto de sinónimos (synsets), siendo estas palabras intercambiables en un contexto.
   b) La integración de WordNet en sistemas de RI permite la expansión de consultas y la mejora de la precisión de los resultados al entender mejor el significado de los términos de búsqueda a través de su contexto semántico y las relaciones entre palabras.
   c) WordNet diferencia claramente entre los significados de palabras según su uso en diferentes contextos, lo que permite aplicaciones avanzadas en desambiguación semántica más allá de los sistemas de recomendación y RI.
   d) Una limitación de WordNet en la representación del conocimiento es su enfoque en el idioma inglés, lo que plantea desafíos en la aplicación global y la interoperabilidad con sistemas de información multilingües.

16. La afirmación que mejor describe la política de amabilidad en los Web Crawlers es:
   a) Los crawlers se diseñan para acceder a sitios web sin restricciones y extraer datos de manera agresiva para su indexación.
   b) La política de amabilidad de los Web Crawlers dicta que los crawlers deben priorizar ciertos tipos de contenido sobre otros, ignorando completamente ciertas páginas web.
   c) La política de amabilidad establece pautas y reglas sobre cómo los crawlers deben interactuar con los sitios web para minimizar la carga del servidor y respetar las directivas de los administradores del sitio.
   d) Los Web Crawlers son libres de recopilar datos de cualquier sitio web sin restricciones, independientemente de la cantidad de tráfico que generen.

17. El concepto de *rank sink* en el algoritmo de PageRank representa páginas web:

a) Con distribuciones uniformes de la puntuación de PageRank.

b) Que tienen una puntuación de PageRank más alta que otras debido a la manipulación de enlaces entrantes y salientes.

c) Con una baja calidad de contenido y una cantidad insuficiente de enlaces salientes, lo que las hace menos relevantes en los resultados de búsqueda.

d) Con un alto número de enlaces salientes que no reciben enlaces entrantes, lo que puede afectar negativamente su puntuación de PageRank.

18. ¿Cuál de las siguientes opciones describe mejor la diferencia clave entre Web Crawling y Web Scraping?

a) Web Crawling se centra en la exploración y recopilación de enlaces de múltiples sitios web, mientras que Web Scraping se enfoca en la extracción específica de datos de páginas web individuales.

b) Web Crawling se realiza utilizando herramientas de automatización como Selenium WebDriver, mientras que Web Scraping se lleva a cabo mediante el análisis de HTML y CSS.

c) Web Scraping implica el análisis de la estructura y el contenido de las páginas web para extraer datos, mientras que Web Crawling se refiere a la descarga y almacenamiento de páginas web completas.

d) Web Scraping es más eficaz para rastrear e indexar contenido web para motores de búsqueda, mientras que Web Crawling se utiliza principalmente para la extracción de datos en proyectos de investigación.

19. En una empresa comercial, se tiene una red donde los nodos corresponden a los empleados y las aristas representan la frecuencia con la que colaboran en las ventas. Se busca mejorar la colaboración entre los empleados para aumentar las ventas totales, por lo que la directiva debe:

a) Utilizar la centralidad de cercanía para identificar los empleados menos cercanos a otros en la red y asignarles tareas individuales para evitar posibles conflictos y desacuerdos en el proceso de colaboración.

b) Utilizar la centralidad de cercanía para identificar los empleados más cercanos a otros en la red y promover la colaboración entre ellos, facilitando así la comunicación y el intercambio de conocimientos para mejorar las ventas.

c) No considerar la centralidad de cercanía de los empleados al diseñar estrategias de colaboración, ya que todos los empleados tienen el mismo acceso a la información y los recursos.

d) No tener en cuenta la centralidad de cercanía de los empleados al diseñar estrategias de colaboración, ya que esto podría introducir complicaciones adicionales en el proceso de trabajo.

20. Un investigador necesita recopilar datos de múltiples sitios web para un estudio académico, pero se enfrenta a varios desafíos al realizar el proceso de Web Scraping de manera ética y legal. ¿Cuál de las siguientes opciones describe mejor uno de los desafíos asociados al proceso de Web Scraping?

a) La necesidad de comprender la estructura del sitio web y su código HTML para extraer los datos correctamente.

b) La disponibilidad limitada de datos en línea que se pueden extraer utilizando técnicas de Web Scraping.

c) La dificultad para encontrar herramientas de Web Scraping gratuitas y fiables.

d) La necesidad de estar montado sobre un scrawler que cumpla con todas las políticas.

Trabajo de Control
Sistemas de Recuperación de Información
Fecha: 15 de julio de 2024
Temario #39

Nombre: _____ Grupo: _____

Seleccione en cada caso la(s) respuesta(s) correcta(s). Considere también no marcar ninguna opción.

1. En el contexto de la representación del conocimiento basada en herencia, ¿qué caracteriza a la herencia cancelable?
   a) La herencia es el resultado del razonamiento no transitivo.
   b) Las conclusiones no están determinadas y dependen del nodo de interés.
   c) Las propiedades heredadas siempre se mantienen y no pueden anularse.
   d) No existe ambigüedad en las conclusiones obtenidas.

2. Se puede afirmar que:
   a) El término "Big Data" se refiere exclusivamente al volumen de datos que una organización maneja, sin tener en cuenta la velocidad, la variedad y la veracidad de los datos.
   b) Los SRI pueden beneficiarse de MapReduce para mejorar la RI relevante.
   c) MapReduce es un modelo de procesamiento distribuido utilizado para trabajar con grandes volúmenes de datos.
   d) Uno de los desafíos en el procesamiento de Big Data es la capacidad de gestionar y analizar datos provenientes de diversas fuentes y en diferentes formatos de manera eficiente.

3. Considerando las dimensiones y desafíos inherentes a Big Data puede afirmarse, tomando en cuenta las características clave y las implicaciones para su procesamiento y análisis, que:
   a) Big Data se caracteriza principalmente por su pequeño volumen y uniformidad, permitiendo un procesamiento eficiente con mínimas adaptaciones de las herramientas de análisis de datos tradicionales.
   b) Big Data no desafía la capacidad de las herramientas tradicionales de procesamiento de datos para capturar, almacenar, gestionar y analizar efectivamente la información, dada la evolución constante de las capacidades computacionales y algoritmos de optimización.
   c) Aunque Big Data puede incluir datos estructurados, su naturaleza se expande al incorporar grandes cantidades de datos no estructurados y semiestructurados, lo que exige el uso de tecnologías especializadas en almacenamiento y procesamiento como Hadoop y sistemas de bases de datos NoSQL.
   d) Además de su complejidad y diversidad, Big Data introduce desafíos significativos en términos de veracidad y variabilidad al requerir métodos avanzados de limpieza y validación de datos para asegurar la integridad del análisis.

4. Al "relajar" el concepto de clique en la detección de comunidades se intenta solucionar:
   a) La imposibilidad de conectar nodos distantes.
   b) La necesidad de datos externos para analizar la red.
   c) La uniformidad de los nodos en términos de grado.
   d) El solapamiento y la complejidad computacional.

5. Se tiene una red de ingredientes donde cada uno representa un nodo y las aristas simbolizan que los ingredientes forman parte de una misma receta. Se busca mejorar la experiencia culinaria mediante la elaboración de combinaciones de ingredientes más interesantes y creativas, para lo cual se debe:
   a) No considerar la centralidad de grado de los ingredientes, ya que todas las combinaciones de ingredientes son igualmente válidas.
   b) Utilizar la centralidad de grado para identificar los ingredientes menos conectados en la red y tomarlos en cuenta para su inclusión en las combinaciones.
   c) No tener en cuenta la centralidad de grado de los ingredientes para no darle mayor importancia a los ingredientes más comunes.
   d) Utilizar la centralidad de grado para identificar los ingredientes más populares en la red y crear combinaciones que incluyan una variedad de ingredientes menos comunes.

6. Un sistema cuenta con la siguiente información:
   - Juan pertenece a la marina.
   - Juan es capellán (sacerdote encargado de una tarea específica fuera de la parroquia).
   - Los infantes de la marina suelen ser bebedores de cerveza.
   - Un capellán no suele ser bebedor de cerveza.
   - Un bebedor de cerveza suele tener sobrepeso.
   - Por lo general, un infante de la marina está en buena condición física.

Si se representa la información en una red de herencia se puede concluir que:

a) No se puede asegurar que Juan sea capellán.

b) No se puede asegurar que Juan esté en buena condición física.

c) Juan tiene sobrepeso.

d) Existen dos razonamientos cancelables.

7. Los Web Crawlers se enfrentan a desafíos constantes. Dentro de ellos se encuentran:

a) La dificultad para generar el contenido dinámico en tiempo real.

b) La modificación del código y la estructura del sitio web.

c) La incapacidad para interpretar correctamente el lenguaje de programación utilizado en el desarrollo de los sitios web.

d) La falta de acceso a la base de datos del servidor web para extraer información actualizada.

8. La Web 1.0 se caracteriza por:

a) Las actualizaciones del contenido de los sitios web se efectúa de forma manual dentro del código HTML

b) Los sitios web se centran en brindar información en lugar de facilitar la colaboración o participación de los usuarios.

c) Uso de comunidades virtuales para popularizar los sitios web de noticias.

d) Los propietarios de los sitios web proporcionan contenido de forma periódica.

9. Un investigador necesita recopilar datos de múltiples sitios web para un estudio académico, pero se enfrenta a varios desafíos al realizar el proceso de Web Scraping de manera ética y legal. ¿Cuál de las siguientes opciones describe mejor uno de los desafíos asociados al proceso de Web Scraping?

a) La necesidad de comprender la estructura del sitio web y su código HTML para extraer los datos correctamente.

b) La disponibilidad limitada de datos en línea que se pueden extraer utilizando técnicas de Web Scraping.

c) La dificultad para encontrar herramientas de Web Scraping gratuitas y fiables.

d) La necesidad de estar montado sobre un scrawler que cumpla con todas las políticas.

10. Los algoritmos para detectar comunidades en una red intentan:

a) Buscar conjuntos donde cada nodo de un mismo conjunto tenga características similares al resto de los nodos del conjunto.

b) Seleccionar aleatoriamente nodos de alto grado y sus vecinos.

c) Buscar subgrafos tal que no incluyan nodos cuya ausencia desconecte al subgrafo.

d) Encontrar grupos donde los nodos pertenecientes a los mismos grupos son cercanos bajo cierta métrica y lejanos con respecto a los nodos de otros grupos.

11. Al realizar la optimización de contenido para SEO debe considerarse:

a) Seleccionar las palabras clave al azar.

b) Utilizar etiquetas de título y meta descripciones únicas y relevantes para cada página.

c) Incluir palabras clave de manera excesiva en el contenido para mejorar el posicionamiento.

d) Crear contenido valioso y original que satisfaga las necesidades de los usuarios.

12. Se puede afirmar que:

a) No es posible identificar subgrupos dentro de una red utilizando análisis de redes.

b) La cantidad de conexiones de un nodo siempre indica su influencia en la red.

c) El tamaño de una red es siempre indicativo de su efectividad en la transmisión de información.

d) Todas las relaciones en una red tienen la misma importancia para el análisis.

13. El propósito de la política de revisitado en los Web Crawlers es:

a) Determinar la frecuencia con la que los crawlers deben volver a visitar una URL específica para mantener la información actualizada en el índice de búsqueda.

b) Establecer reglas sobre el tiempo máximo que los crawlers pueden pasar en un sitio web durante cada visita para evitar sobrecargar los servidores.

c) Limitar el acceso de los crawlers a ciertos servidores luego de visitar las páginas alojadas en estos.

d) Definir la prioridad de rastreo de las URLs, determinando el orden en que los crawlers visitan y procesan cada página web.

14. ¿Qué estrategia utilizan los Web Crawlers para asegurar un rastreo eficiente y respetuoso de los recursos de los sitios web?

a) Ignorar completamente el archivo `robots.txt` de los sitios web.

b) Visitar y rastrear todos los enlaces de una página web simultáneamente.

c) Extraer únicamente contenido multimedia para reducir la carga en los servidores web.

d) Seguir las directrices del archivo `robots.txt` y aplicar un retraso entre las solicitudes.

15. Si se tiene el conjunto de páginas interconectadas $\{A \to B, C; B \to C; C \to A\}$, entonces la página con valor más alto de PageRank es:
    a) A y B.
    b) A.
    c) C.
    d) B.

16. En una biblioteca digital se necesita implementar un sistema que permita a los usuarios encontrar libros y artículos científicos de forma rápida y precisa. Los documentos están en diversos formatos, incluyendo PDF, EPUB y HTML. Se requiere seleccionar un algoritmo de indexación adecuado para el sistema, por lo que el programador designado para la implementación debe considerar:
    a) La velocidad de indexación y recuperación de datos.
    b) La capacidad para manejar documentos en diferentes formatos.
    c) La complejidad del algoritmo en términos de implementación y mantenimiento.
    d) La capacidad del algoritmo para procesar imágenes incrustadas o referenciadas en los ficheros.

17. Una plataforma de comercio electrónico desea mejorar la experiencia del usuario al permitir una navegación más personalizada y contextualizada. Actualmente los usuarios tienen dificultades para encontrar productos relevantes debido a la gran cantidad de opciones disponibles. La empresa está interesada en implementar características de la Web 2.5 y la Web Semántica para abordar este problema. ¿Qué características podrían ayudar para ofrecer una navegación más personalizada y contextualizada?
    a) La implementación de ontologías y metadatos para enriquecer la descripción de productos y mejorar la precisión de las recomendaciones.
    b) La optimización de la velocidad de carga del sitio web para mejorar la experiencia del usuario y reducir el abandono del carrito de compra.
    c) La integración de redes sociales para permitir la recomendación de productos basada en las preferencias de amigos y contactos.
    d) La incorporación de los productos en tendencia en el mercado.

18. Una empresa de comercio electrónico necesita mejorar su motor de búsqueda para proporcionar resultados más relevantes a sus usuarios. Actualmente, los resultados de la búsqueda no son precisos y los usuarios a menudo encuentran dificultades para encontrar productos específicos. La empresa está considerando implementar una indexación por conceptos para mejorar la relevancia de los resultados de búsqueda. ¿Qué beneficios podría provocar este cambio?
    a) Permite adaptarse fácilmente a cambios en el vocabulario y la terminología utilizada en las descripciones de los productos.
    b) Ayuda a identificar automáticamente relaciones entre productos, lo que puede mejorar las recomendaciones personalizadas a los usuarios.
    c) Facilita la visualización de productos al agruparlos por categorías o características comunes.
    d) Mejora la precisión en la búsqueda de productos relacionados, incluso cuando no coinciden exactamente con los términos de búsqueda del usuario.

19. Analizar una red permite:
    a) Detectar posibles tendencias antes de que se conviertan en tendencia.
    b) Obtener predicciones exactas de eventos futuros en mercados financieros.
    c) Evaluar la calidad del contenido de un sitio web.
    d) Encontrar nodos "sensibles" o críticos para la red.

20. El concepto de *rank sink* en el algoritmo de PageRank representa páginas web:
    a) Con distribuciones uniformes de la puntuación de PageRank.
    b) Que tienen una puntuación de PageRank más alta que otras debido a la manipulación de enlaces entrantes y salientes.
    c) Con una baja calidad de contenido y una cantidad insuficiente de enlaces salientes, lo que las hace menos relevantes en los resultados de búsqueda.
    d) Con un alto número de enlaces salientes que no reciben enlaces entrantes, lo que puede afectar negativamente su puntuación de PageRank.

Trabajo de Control
Sistemas de Recuperación de Información
Fecha: 15 de julio de 2024
Temario #40

Nombre: _____ Grupo: _____

Seleccione en cada caso la(s) respuesta(s) correcta(s). Considere también no marcar ninguna opción.

1. En el diseño de un videojuego de roles (RPG) en el que los personajes tienen atributos como salud, fuerza y velocidad, ¿cuál de las siguientes opciones representa mejor una implementación de la representación del conocimiento orientado a objetos?
   a) Cada personaje se representa como una lista de cadenas de texto que describen sus características físicas y habilidades.
   b) Cada personaje se representa como una función que calcula sus atributos en función de su nivel y experiencia.
   c) Cada personaje se representa como una matriz de números que almacena sus valores de atributos.
   d) Cada personaje se representa como un objeto con propiedades como salud, fuerza y velocidad, y métodos para modificar y consultar estos valores.

2. Analizar una red permite:
   a) Detectar posibles tendencias antes de que se conviertan en tendencia.
   b) Obtener predicciones exactas de eventos futuros en mercados financieros.
   c) Evaluar la calidad del contenido de un sitio web.
   d) Encontrar nodos "sensibles" o críticos para la red.

3. Dentro del ecosistema de Hadoop, el HDFS se caracteriza por:
   a) El modelo de acceso y de escritura de datos en tiempo real.
   b) La tolerancia a fallos mediante la replicación de datos.
   c) El almacenamiento exclusivo para archivos de texto.
   d) La capacidad ilimitada de almacenamiento.

4. Sobre WordNet y su aplicación en el procesamiento del lenguaje natural, se puede afirmar que:
   a) El diseño de WordNet facilita su integración en aplicaciones multilingües de NLP, aunque su desarrollo original se centró en el inglés.
   b) Aunque WordNet es una herramienta valiosa en el NLP, su estructura no incluye información sobre la frecuencia de uso de las palabras en el lenguaje natural.
   c) Los synsets facilitan la identificación de relaciones semánticas entre palabras, como la hiperonimia y la meronimia, enriqueciendo tareas de NLP.
   d) WordNet proporciona una base para la desambiguación semántica al agrupar palabras con significados similares en synsets.

5. En un SRI la indexación:
   a) Consiste en asociar un identificador único a cada dato almacenado en el sistema.
   b) Mejora la experiencia del usuario.
   c) Permite la organización y la categorización de la información.
   d) Facilita la RI relevante.

6. Al realizar la optimización de contenido para SEO debe considerarse:
   a) Seleccionar las palabras clave al azar.
   b) Utilizar etiquetas de título y meta descripciones únicas y relevantes para cada página.
   c) Incluir palabras clave de manera excesiva en el contenido para mejorar el posicionamiento.
   d) Crear contenido valioso y original que satisfaga las necesidades de los usuarios.

7. El algoritmo de PageRank converge si:
   a) La norma de la diferencia entre los vectores es menor a un umbral predefinido.
   b) Finaliza la ejecución cuando el número de iteraciones excede un máximo de iteraciones previamente definido.
   c) Se define un factor de normalización en la fórmula de la función.
   d) El algoritmo no se implementa de forma iterativa.

8. Una buena práctica de SEO para mejorar el posicionamiento de un sitio web en los motores de búsqueda es:
   a) Obtener enlaces de otros sitios web relevantes y de calidad que apunten al sitio.
   b) Copiar contenido directamente de otros sitios web populares para aumentar la cantidad de páginas indexadas.
   c) Utilizar técnicas de encubrimiento para mostrar contenido diferente a lo indexado por los motores de búsqueda.

d) Llenar el contenido con palabras clave irrelevantes para aumentar la densidad de palabras clave.

9. La transición de la Web 1.0 a la Web 2.0 se caracterizó principalmente por:
   a) El aumento en la velocidad de conexión a internet, que permitió una mejor calidad de las páginas web.
   b) La disminución de la importancia de los motores de búsqueda en la navegación web.
   c) El cambio de páginas web estáticas a dinámicas, permitiendo la interacción del usuario y la generación de contenido.
   d) La reducción en el uso de HTML y CSS en el desarrollo de sitios web.

10. En una red de transporte donde cada nodo es una parada de autobús y las aristas representan si existe un carro que pasa por ambos sitios, ¿qué puede mejorar el sistema de transporte?
    a) No considerar la centralidad de intermediación de las estaciones de transporte público al planificar rutas y horarios, ya que no tiene impacto en las conexiones entre las paradas.
    b) Utilizar el grafo inducido de los nodos con mayor valor en la centralidad de grado para aplicar la centralidad de intermediación con el propósito de reforzar las paradas con mayor tráfico.
    c) Utilizar la centralidad de intermediación para identificar las paradas de transferencia clave y establecer nuevas conexiones entre diferentes líneas de transporte público.
    d) No implementar sistemas de información en tiempo real para los usuarios, ya que pueden aumentar la carga de los trabajadores.

11. La política de ordenación de URLs en los Web Crawlers tiene como aspecto fundamental:
    a) Limitar el acceso de los crawlers a ciertas secciones de un sitio web, evitando el rastreo de URLs consideradas menos importantes o sensibles.
    b) Definir la estructura de la URL de destino, asegurando que estén ordenadas alfabéticamente para facilitar la navegación y la indexación.
    c) Establecer la prioridad de rastreo de las URLs, determinando el orden en que los crawlers visitan y procesan cada página web.
    d) Determinar la forma en que los crawlers asignan un valor de relevancia a cada URL para clasificarlas en el índice de búsqueda.

12. Considerando las prácticas éticas y legales en el Web Scraping, se puede asegurar que:
    a) Web Scraping requiere considerar las políticas de `robots.txt` del sitio web objetivo.
    b) Es importante revisar y respetar los términos de servicio del sitio web, así como las leyes aplicables de protección de datos y derechos de autor, antes de realizar Web Scraping.
    c) Web Scraping sobre datos personales sin consentimiento es generalmente aceptado si los datos se utilizan con fines de investigación.
    d) La extracción de datos mediante Web Scraping siempre es legal, independientemente de las leyes locales sobre derechos de autor y privacidad de datos definidos en los sitios web.

13. Un sistema cuenta con la siguiente información:
    - Juan pertenece a la marina.
    - Juan es capellán (sacerdote encargado de una tarea específica fuera de la parroquia).
    - Los infantes de la marina suelen ser bebedores de cerveza.
    - Un capellán no suele ser bebedor de cerveza.
    - Un bebedor de cerveza suele tener sobrepeso.
    - Por lo general, un infante de la marina está en buena condición física.

    Si se representa la información en una red de herencia se puede concluir que:
    a) No se puede asegurar que Juan sea capellán.
    b) No se puede asegurar que Juan esté en buena condición física.
    c) Juan tiene sobrepeso.
    d) Existen dos razonamientos cancelables.

14. En la RI en el contexto de Big Data, se puede asegurar que:
    a) La indexación distribuida divide los datos en múltiples fragmentos que se almacenan en varios nodos para permitir búsquedas paralelas y mejorar la escalabilidad.
    b) La indexación distribuida no ofrece ventajas en términos de rendimiento y escalabilidad en comparación con la indexación centralizada.
    c) La indexación distribuida es una técnica obsoleta.
    d) La indexación distribuida almacena todos los datos en un solo servidor para facilitar su acceso y búsqueda.

15. El uso de Hadoop y MapReduce en el contexto de la RI tiene como objetivo:
    a) Automatizar el mantenimiento de sistemas de bases de datos.
    b) Facilitar el análisis en tiempo real de datos de redes sociales.
    c) Mejorar la eficiencia energética en centros de datos.
    d) Procesar y analizar grandes conjuntos de datos para la RI.

16. En el algoritmo de Indexación Basada en Clasificación Bloqueada (BSBI), ¿cuál es el paso final para crear un índice invertido para la colección completa de los datos?

    a) Eliminar los términos duplicados de los índices.

    b) Indexar cada bloque de forma independiente.

    c) Fusionar los índices invertidos de cada bloque.

    d) Dividir la colección de datos en bloques de tamaño fijo.

17. Con respecto a WordNet y su contribución a las ontologías y la representación del conocimiento en los sistemas de RI, puede afirmarse que:

    a) En WordNet, una palabra está asociada a un conjunto de sinónimos (synsets), siendo estas palabras intercambiables en un contexto.

    b) La integración de WordNet en sistemas de RI permite la expansión de consultas y la mejora de la precisión de los resultados al entender mejor el significado de los términos de búsqueda a través de su contexto semántico y las relaciones entre palabras.

    c) WordNet diferencia claramente entre los significados de palabras según su uso en diferentes contextos, lo que permite aplicaciones avanzadas en desambiguación semántica más allá de los sistemas de recomendación y RI.

    d) Una limitación de WordNet en la representación del conocimiento es su enfoque en el idioma inglés, lo que plantea desafíos en la aplicación global y la interoperabilidad con sistemas de información multilingües.

18. Se puede afirmar que:

    a) El término "Big Data" se refiere exclusivamente al volumen de datos que una organización maneja, sin tener en cuenta la velocidad, la variedad y la veracidad de los datos.

    b) Los SRI pueden beneficiarse de MapReduce para mejorar la RI relevante.

    c) MapReduce es un modelo de procesamiento distribuido utilizado para trabajar con grandes volúmenes de datos.

    d) Uno de los desafíos en el procesamiento de Big Data es la capacidad de gestionar y analizar datos provenientes de diversas fuentes y en diferentes formatos de manera eficiente.

19. Si se tiene el conjunto de páginas interconectadas $\{A \rightarrow B, C; B \rightarrow C; C \rightarrow A\}$, entonces la página con valor más alto de PageRank es:

    a) A y B.

    b) A.

    c) C.

    d) B.

20. La centralidad de intermediación de un nodo indica:

    a) La cantidad de nodos vecinos directos.

    b) La resistencia del nodo a fallos.

    c) El número total de conexiones entrantes y salientes.

    d) La frecuencia con la que un nodo actúa como puente en el camino más corto entre otros dos nodos.

Trabajo de Control
Sistemas de Recuperación de Información
Fecha: 15 de julio de 2024
Temario #41

Nombre: _____ Grupo: _____

Seleccione en cada caso la(s) respuesta(s) correcta(s). Considere también no marcar ninguna opción.

1. Sobre el algoritmo de PageRank, visto en clase, se puede afirmar que:
   a) Evalúa la importancia de un sitio web en función de la calidad y cantidad de enlaces entrantes que recibe de otros sitios web.
   b) Solo tiene en cuenta el contenido en un sitio web para determinar su relevancia en los resultados de búsqueda.
   c) Asigna una puntuación alta a los sitios web que tienen un gran número de enlaces entrantes sin tener en cuenta la calidad de esos enlaces.
   d) Asigna una puntuación baja a los sitios web que contienen muchos enlaces salientes, ya que indica una falta de relevancia.

2. En una biblioteca digital se necesita implementar un sistema que permita a los usuarios encontrar libros y artículos científicos de forma rápida y precisa. Los documentos están en diversos formatos, incluyendo PDF, EPUB y HTML. Se requiere seleccionar un algoritmo de indexación adecuado para el sistema, por lo que el programador designado para la implementación debe considerar:
   a) La velocidad de indexación y recuperación de datos.
   b) La capacidad para manejar documentos en diferentes formatos.
   c) La complejidad del algoritmo en términos de implementación y mantenimiento.
   d) La capacidad del algoritmo para procesar imágenes incrustadas o referenciadas en los ficheros.

3. Considerando las dimensiones y desafíos inherentes a Big Data puede afirmarse, tomando en cuenta las características clave y las implicaciones para su procesamiento y análisis, que:
   a) Big Data se caracteriza principalmente por su pequeño volumen y uniformidad, permitiendo un procesamiento eficiente con mínimas adaptaciones de las herramientas de análisis de datos tradicionales.
   b) Big Data no desafía la capacidad de las herramientas tradicionales de procesamiento de datos para capturar, almacenar, gestionar y analizar efectivamente la información, dada la evolución constante de las capacidades computacionales y algoritmos de optimización.
   c) Aunque Big Data puede incluir datos estructurados, su naturaleza se expande al incorporar grandes cantidades de datos no estructurados y semiestructurados, lo que exige el uso de tecnologías especializadas en almacenamiento y procesamiento como Hadoop y sistemas de bases de datos NoSQL.
   d) Además de su complejidad y diversidad, Big Data introduce desafíos significativos en términos de veracidad y variabilidad al requerir métodos avanzados de limpieza y validación de datos para asegurar la integridad del análisis.

4. ¿Qué es un "Uniform Resource Locator (URL) Frontier" en el contexto de Web Crawling?
   a) Un protocolo que define cómo se deben formatear las URLs para el crawling.
   b) Una lista prioritaria de URLs que aún no han sido visitadas por el crawler.
   c) Una técnica para filtrar URLs irrelevantes y mejorar la eficiencia del crawling.
   d) Una base de datos que almacena URLs únicas identificadas como recursos en la web.

5. La política de ordenación de URLs en los Web Crawlers tiene como aspecto fundamental:
   a) Limitar el acceso de los crawlers a ciertas secciones de un sitio web, evitando el rastreo de URLs consideradas menos importantes o sensibles.
   b) Definir la estructura de la URL de destino, asegurando que estén ordenadas alfabéticamente para facilitar la navegación y la indexación.
   c) Establecer la prioridad de rastreo de las URLs, determinando el orden en que los crawlers visitan y procesan cada página web.
   d) Determinar la forma en que los crawlers asignan un valor de relevancia a cada URL para clasificarlas en el índice de búsqueda.

6. La afirmación que mejor refleja el principio subyacente de PageRank, considerando su importancia en la clasificación de los sitios web, es:
   a) PageRank valora más la cantidad de enlaces entrantes a una página web, independientemente de la calidad o relevancia de estos enlaces.

b) La efectividad de PageRank se basa exclusivamente en el análisis de las palabras clave contenidas en los enlaces entrantes, sin considerar la estructura de enlace de la Web.

c) El algoritmo de PageRank considera tanto la cantidad como la calidad de los enlaces entrantes, asignando mayor valor a los enlaces provenientes de sitios web considerados como "importantes".

d) PageRank opera bajo el supuesto de que los enlaces entrantes y salientes tienen el mismo impacto en la valoración de la relevancia de una página web.

7. En el diseño de un videojuego de roles (RPG) en el que los personajes tienen atributos como salud, fuerza y velocidad, ¿cuál de las siguientes opciones representa mejor una implementación de la representación del conocimiento orientado a objetos?

   a) Cada personaje se representa como una lista de cadenas de texto que describen sus características físicas y habilidades.

   b) Cada personaje se representa como una función que calcula sus atributos en función de su nivel y experiencia.

   c) Cada personaje se representa como una matriz de números que almacena sus valores de atributos.

   d) Cada personaje se representa como un objeto con propiedades como salud, fuerza y velocidad, y métodos para modificar y consultar estos valores.

8. Una plataforma de comercio electrónico desea mejorar la experiencia del usuario al permitir una navegación más personalizada y contextualizada. Actualmente los usuarios tienen dificultades para encontrar productos relevantes debido a la gran cantidad de opciones disponibles. La empresa está interesada en implementar características de la Web 2.5 y la Web Semántica para abordar este problema. ¿Qué características podrían ayudar para ofrecer una navegación más personalizada y contextualizada?

   a) La implementación de ontologías y metadatos para enriquecer la descripción de productos y mejorar la precisión de las recomendaciones.

   b) La optimización de la velocidad de carga del sitio web para mejorar la experiencia del usuario y reducir el abandono del carrito de compra.

   c) La integración de redes sociales para permitir la recomendación de productos basada en las preferencias de amigos y contactos.

   d) La incorporación de los productos en tendencia en el mercado.

9. En la RI en el contexto de Big Data, se puede asegurar que:

   a) La indexación distribuida divide los datos en múltiples fragmentos que se almacenan en varios nodos para permitir búsquedas paralelas y mejorar la escalabilidad.

   b) La indexación distribuida no ofrece ventajas en términos de rendimiento y escalabilidad en comparación con la indexación centralizada.

   c) La indexación distribuida es una técnica obsoleta.

   d) La indexación distribuida almacena todos los datos en un solo servidor para facilitar su acceso y búsqueda.

10. En una empresa comercial, se tiene una red donde los nodos corresponden a los empleados y las aristas representan la frecuencia con la que colaboran en las ventas. Se busca mejorar la colaboración entre los empleados para aumentar las ventas totales, por lo que la directiva debe:

    a) Utilizar la centralidad de cercanía para identificar los empleados menos cercanos a otros en la red y asignarles tareas individuales para evitar posibles conflictos y desacuerdos en el proceso de colaboración.

    b) Utilizar la centralidad de cercanía para identificar los empleados más cercanos a otros en la red y promover la colaboración entre ellos, facilitando así la comunicación y el intercambio de conocimientos para mejorar las ventas.

    c) No considerar la centralidad de cercanía de los empleados al diseñar estrategias de colaboración, ya que todos los empleados tienen el mismo acceso a la información y los recursos.

    d) No tener en cuenta la centralidad de cercanía de los empleados al diseñar estrategias de colaboración, ya que esto podría introducir complicaciones adicionales en el proceso de trabajo.

11. Considerando las prácticas éticas y legales en el Web Scraping, se puede asegurar que:

    a) Web Scraping requiere considerar las políticas de `robots.txt` del sitio web objetivo.

    b) Es importante revisar y respetar los términos de servicio del sitio web, así como las leyes aplicables de protección de datos y derechos de autor, antes de realizar Web Scraping.

    c) Web Scraping sobre datos personales sin consentimiento es generalmente aceptado si los datos se utilizan con fines de investigación.

    d) La extracción de datos mediante Web Scraping siempre es legal, independientemente de las leyes locales sobre derechos de autor y privacidad de datos definidos en los sitios web.

12. La Web 3 se conoce como:

    a) Internet de las cosas.

b) Web de solo lectura.

c) Web semántica.

d) Web de escritura-lectura.

13. En un sistema de control de tráfico urbano basado en reglas, ¿cuál de las siguientes reglas sería más efectiva para manejar situaciones de congestión en una intersección?

a) Si hay pocos vehículos en la intersección, reducir el tiempo de los semáforos en verde.

b) Si hay muchos vehículos en la intersección, aumentar el tiempo de los semáforos en verde.

c) Si hay un vehículo de emergencia en la intersección, detener todos los demás vehículos.

d) Si hay muchos peatones cruzando la intersección, reducir el tiempo de los semáforos en rojo.

14. La centralidad de intermediación de un nodo indica:

a) La cantidad de nodos vecinos directos.

b) La resistencia del nodo a fallos.

c) El número total de conexiones entrantes y salientes.

d) La frecuencia con la que un nodo actúa como puente en el camino más corto entre otros dos nodos.

15. En el modelo de representación del conocimiento basado en herencia se puede asegurar que:

a) La cantidad de padres que puede tener un nodo no es mayor que 4.

b) Las conclusiones pueden ser canceladas si el grafo es ambiguo.

c) Solo se usa en entornos referentes a la biología.

d) El razonamiento deducido está respaldado por al menos un camino dentro del grafo.

16. La transición de la Web 1.0 a la Web 2.0 se caracterizó principalmente por:

a) El aumento en la velocidad de conexión a internet, que permitió una mejor calidad de las páginas web.

b) La disminución de la importancia de los motores de búsqueda en la navegación web.

c) El cambio de páginas web estáticas a dinámicas, permitiendo la interacción del usuario y la generación de contenido.

d) La reducción en el uso de HTML y CSS en el desarrollo de sitios web.

17. El algoritmo de PageRank converge si:

a) La norma de la diferencia entre los vectores es menor a un umbral predefinido.

b) Finaliza la ejecución cuando el número de iteraciones excede un máximo de iteraciones previamente definido.

c) Se define un factor de normalización en la fórmula de la función.

d) El algoritmo no se implementa de forma iterativa.

18. El algoritmo de PageRank puede describirse como un procedimiento utilizado para:

a) Determinar la velocidad de carga de una página web en un navegador.

b) Calcular la relevancia de una página web en función de la cantidad y calidad de los enlaces que apuntan hacia ella.

c) Prevenir el spam y el contenido no deseado en las páginas web.

d) Clasificar las páginas web en función de su edad y autoridad.

19. En un sistema donde el conocimiento está orientado a objetos se puede asegurar que:

a) Se enfatiza la atención a la información de la cual se extrajo el conocimiento.

b) Los marcos y las bandas son las estructuras utilizadas para la representación del modelo.

c) No es posible definir especializaciones de los objetos de la vida real dentro del sistema.

d) Las funciones de agregación dificultan poder establecer relaciones entre los objetos.

20. ¿Qué algoritmos permiten obtener información de una red?

a) Índices de centralidad.

b) Detección de comunidades.

c) Hypertext Induced Topic Selection (HITS).

d) PageRank.

Trabajo de Control
Sistemas de Recuperación de Información
Fecha: 15 de julio de 2024
Temario #42

Nombre: _____ Grupo: _____

Seleccione en cada caso la(s) respuesta(s) correcta(s). Considere también no marcar ninguna opción.

1. En el algoritmo de Indexación Basada en Clasificación Bloqueada (BSBI), ¿cuál es el paso final para crear un índice invertido para la colección completa de los datos?
   a) Eliminar los términos duplicados de los índices.
   b) Indexar cada bloque de forma independiente.
   c) Fusionar los índices invertidos de cada bloque.
   d) Dividir la colección de datos en bloques de tamaño fijo.

2. Dentro del análisis de redes, la centralidad de grado mide la importancia de un nodo basándose en:
   a) El grado del nodo.
   b) La cantidad de veces que aparece el nodo en el camino mínimo entre cualquier par de nodos.
   c) La cantidad de vecinos del nodo.
   d) El número de aristas que posee el nodo.

3. La computación evolutiva:
   a) No es aplicable en la RI debido a la complejidad de los algoritmos evolutivos.
   b) Solo puede manejar conjuntos de datos pequeños y no es escalable a grandes volúmenes de datos.
   c) Utiliza algoritmos para buscar soluciones óptimas en grandes espacios de búsqueda, lo que la hace adecuada para problemas de optimización en la RI.
   d) Es útil solo para problemas de clasificación de documentos y no para otras tareas de RI en general.

4. El posicionamiento de un sitio web en los motores de búsqueda puede ser afectado por:
   a) Utilizar técnicas de encubrimiento para mostrar contenido diferente a lo indexado por los motores de búsqueda.
   b) Contenido sin valor en el sitio web.
   c) Mantener una estructura de URL clara y coherente.
   d) Obtener enlaces de sitios web irrelevantes y de baja calidad.

5. ¿Qué es un "Uniform Resource Locator (URL) Frontier" en el contexto de Web Crawling?
   a) Un protocolo que define cómo se deben formatear las URLs para el crawling.
   b) Una lista prioritaria de URLs que aún no han sido visitadas por el crawler.
   c) Una técnica para filtrar URLs irrelevantes y mejorar la eficiencia del crawling.
   d) Una base de datos que almacena URLs únicas identificadas como recursos en la web.

6. Los algoritmos para detectar comunidades en una red intentan:
   a) Buscar conjuntos donde cada nodo de un mismo conjunto tenga características similares al resto de los nodos del conjunto.
   b) Seleccionar aleatoriamente nodos de alto grado y sus vecinos.
   c) Buscar subgrafos tal que no incluyan nodos cuya ausencia desconecte al subgrafo.
   d) Encontrar grupos donde los nodos pertenecientes a los mismos grupos son cercanos bajo cierta métrica y lejanos con respecto a los nodos de otros grupos.

7. El algoritmo de PageRank converge si:
   a) La norma de la diferencia entre los vectores es menor a un umbral predefinido.
   b) Finaliza la ejecución cuando el número de iteraciones excede un máximo de iteraciones previamente definido.
   c) Se define un factor de normalización en la fórmula de la función.
   d) El algoritmo no se implementa de forma iterativa.

8. Implementar índices invertidos en un SRI asegura:
   a) Acelerar el proceso de búsqueda al permitir búsquedas directas por contenido en lugar de por título.
   b) Reducir la cantidad de espacio de almacenamiento necesario al comprimir los datos de los documentos.
   c) Facilitar la búsqueda sobre los datos que contienen términos específicos al mantener una lista de datos para cada término único.
   d) Incrementar la seguridad de los datos almacenados al dificultar el acceso directo a la información sin el índice correcto.

9. Para la RI, el análisis de las redes puede:

a) Ayudar a identificar grupos de interés.

b) Indicar la importancia de una entidad en la transmisión de la información.

c) Ayudar a comprender la conectividad y la accesibilidad entre las entidades.

d) Revelar patrones de influencia dentro de una comunidad.

10. Se puede afirmar que:

a) El término "Big Data" se refiere exclusivamente al volumen de datos que una organización maneja, sin tener en cuenta la velocidad, la variedad y la veracidad de los datos.

b) Los SRI pueden beneficiarse de MapReduce para mejorar la RI relevante.

c) MapReduce es un modelo de procesamiento distribuido utilizado para trabajar con grandes volúmenes de datos.

d) Uno de los desafíos en el procesamiento de Big Data es la capacidad de gestionar y analizar datos provenientes de diversas fuentes y en diferentes formatos de manera eficiente.

11. ¿Qué estrategia utilizan los Web Crawlers para asegurar un rastreo eficiente y respetuoso de los recursos de los sitios web?

a) Ignorar completamente el archivo `robots.txt` de los sitios web.

b) Visitar y rastrear todos los enlaces de una página web simultáneamente.

c) Extraer únicamente contenido multimedia para reducir la carga en los servidores web.

d) Seguir las directrices del archivo `robots.txt` y aplicar un retraso entre las solicitudes.

12. Un investigador necesita recopilar datos de múltiples sitios web para un estudio académico, pero se enfrenta a varios desafíos al realizar el proceso de Web Scraping de manera ética y legal. ¿Cuál de las siguientes opciones describe mejor uno de los desafíos asociados al proceso de Web Scraping?

a) La necesidad de comprender la estructura del sitio web y su código HTML para extraer los datos correctamente.

b) La disponibilidad limitada de datos en línea que se pueden extraer utilizando técnicas de Web Scraping.

c) La dificultad para encontrar herramientas de Web Scraping gratuitas y fiables.

d) La necesidad de estar montado sobre un scrawler que cumpla con todas las políticas.

13. En una red de transporte donde cada nodo es una parada de autobús y las aristas representan si existe un carro que pasa por ambos sitios, ¿qué puede mejorar el sistema de transporte?

a) No considerar la centralidad de intermediación de las estaciones de transporte público al planificar rutas y horarios, ya que no tiene impacto en las conexiones entre las paradas.

b) Utilizar el grafo inducido de los nodos con mayor valor en la centralidad de grado para aplicar la centralidad de intermediación con el propósito de reforzar las paradas con mayor tráfico.

c) Utilizar la centralidad de intermediación para identificar las paradas de transferencia clave y establecer nuevas conexiones entre diferentes líneas de transporte público.

d) No implementar sistemas de información en tiempo real para los usuarios, ya que pueden aumentar la carga de los trabajadores.

14. En el diseño de un videojuego de roles (RPG) en el que los personajes tienen atributos como salud, fuerza y velocidad, ¿cuál de las siguientes opciones representa mejor una implementación de la representación del conocimiento orientado a objetos?

a) Cada personaje se representa como una lista de cadenas de texto que describen sus características físicas y habilidades.

b) Cada personaje se representa como una función que calcula sus atributos en función de su nivel y experiencia.

c) Cada personaje se representa como una matriz de números que almacena sus valores de atributos.

d) Cada personaje se representa como un objeto con propiedades como salud, fuerza y velocidad, y métodos para modificar y consultar estos valores.

15. Una buena práctica de SEO para mejorar el posicionamiento de un sitio web en los motores de búsqueda es:

a) Obtener enlaces de otros sitios web relevantes y de calidad que apunten al sitio.

b) Copiar contenido directamente de otros sitios web populares para aumentar la cantidad de páginas indexadas.

c) Utilizar técnicas de encubrimiento para mostrar contenido diferente a lo indexado por los motores de búsqueda.

d) Llenar el contenido con palabras clave irrelevantes para aumentar la densidad de palabras clave.

16. Analizar una red permite:

a) Detectar posibles tendencias antes de que se conviertan en tendencia.

b) Obtener predicciones exactas de eventos futuros en mercados financieros.

c) Evaluar la calidad del contenido de un sitio web.

d) Encontrar nodos "sensibles" o críticos para la red.

17. Para transformar el contenido no estructurado de las páginas web en datos estructurados, el Web Scraping:

a) Extrae información basada en patrones de HTML/CSS.

b) Analiza los protocolos de red.

c) Interpreta el código JavaScript en tiempo real.

d) Convierte de forma automática imágenes a texto.

18. En la RI en el contexto de Big Data, se puede asegurar que:

a) El almacenamiento distribuido centraliza todos los datos en un único servidor para facilitar su gestión y mantenimiento.

b) El almacenamiento distribuido ofrece ventajas significativas en términos de escalabilidad y rendimiento en comparación con el almacenamiento centralizado.

c) El almacenamiento distribuido reparte los datos en múltiples servidores para mejorar la disponibilidad y la redundancia del sistema.

d) El almacenamiento distribuido es una técnica obsoleta.

19. El componente responsable de la gestión de recursos y planificación de tareas en Hadoop es:

a) MapReduce.

b) Hadoop Common.

c) HDFS.

d) YARN.

20. Se puede afirmar que:

a) No es posible identificar subgrupos dentro de una red utilizando análisis de redes.

b) La cantidad de conexiones de un nodo siempre indica su influencia en la red.

c) El tamaño de una red es siempre indicativo de su efectividad en la transmisión de información.

d) Todas las relaciones en una red tienen la misma importancia para el análisis.

Trabajo de Control
Sistemas de Recuperación de Información
Fecha: 15 de julio de 2024
Temario #43

Nombre: _____ Grupo: _____

Seleccione en cada caso la(s) respuesta(s) correcta(s). Considere también no marcar ninguna opción.

1. En una red de transporte donde cada nodo es una parada de autobús y las aristas representan si existe un carro que pasa por ambos sitios, ¿qué puede mejorar el sistema de transporte?
   a) No considerar la centralidad de intermediación de las estaciones de transporte público al planificar rutas y horarios, ya que no tiene impacto en las conexiones entre las paradas.
   b) Utilizar el grafo inducido de los nodos con mayor valor en la centralidad de grado para aplicar la centralidad de intermediación con el propósito de reforzar las paradas con mayor tráfico.
   c) Utilizar la centralidad de intermediación para identificar las paradas de transferencia clave y establecer nuevas conexiones entre diferentes líneas de transporte público.
   d) No implementar sistemas de información en tiempo real para los usuarios, ya que pueden aumentar la carga de los trabajadores.

2. La Web 3 se conoce como:
   a) Internet de las cosas.
   b) Web de solo lectura.
   c) Web semántica.
   d) Web de escritura-lectura.

3. En el contexto de la representación del conocimiento basada en herencia, ¿qué caracteriza a la herencia cancelable?
   a) La herencia es el resultado del razonamiento no transitivo.
   b) Las conclusiones no están determinadas y dependen del nodo de interés.
   c) Las propiedades heredadas siempre se mantienen y no pueden anularse.
   d) No existe ambigüedad en las conclusiones obtenidas.

4. En el contexto de la RI en redes, se puede afirmar que:
   a) La detección de comunidades no es relevante para la RI en redes, ya que se centra únicamente en la estructura de la red sin considerar el contenido.
   b) La detección de comunidades en una red siempre produce resultados objetivos y consistentes independientemente del algoritmo utilizado.
   c) La detección de comunidades ayuda a identificar grupos de nodos altamente conectados entre sí, lo que puede ser útil para comprender la estructura y el contenido de la red.
   d) La detección de comunidades solo se aplica a redes pequeñas y simples, no a redes grandes y complejas.

5. El concepto de *rank sink* en el algoritmo de PageRank representa páginas web:
   a) Con distribuciones uniformes de la puntuación de PageRank.
   b) Que tienen una puntuación de PageRank más alta que otras debido a la manipulación de enlaces entrantes y salientes.
   c) Con una baja calidad de contenido y una cantidad insuficiente de enlaces salientes, lo que las hace menos relevantes en los resultados de búsqueda.
   d) Con un alto número de enlaces salientes que no reciben enlaces entrantes, lo que puede afectar negativamente su puntuación de PageRank.

6. Referente al Web Crawler puede afirmarse que:
   a) Los hipervínculos encontrados en cada sitio web que no pertenecen al dominio donde fueron encontrados se desechan, puesto que no expande el conjunto de URLs sin visitar.
   b) Las páginas visitadas no se procesan nunca más.
   c) No necesita de un conjunto inicial de URLs para recorrer la Web.
   d) No tiene como objetivo indexar y recopilar información de diferentes sitios web.

7. La computación evolutiva:
   a) No es aplicable en la RI debido a la complejidad de los algoritmos evolutivos.
   b) Solo puede manejar conjuntos de datos pequeños y no es escalable a grandes volúmenes de datos.
   c) Utiliza algoritmos para buscar soluciones óptimas en grandes espacios de búsqueda, lo que la hace adecuada para problemas de optimización en la RI.
   d) Es útil solo para problemas de clasificación de documentos y no para otras tareas de RI en general.

8. El propósito de la política de revisitado en los Web Crawlers es:

a) Determinar la frecuencia con la que los crawlers deben volver a visitar una URL específica para mantener la información actualizada en el índice de búsqueda.

b) Establecer reglas sobre el tiempo máximo que los crawlers pueden pasar en un sitio web durante cada visita para evitar sobrecargar los servidores.

c) Limitar el acceso de los crawlers a ciertos servidores luego de visitar las páginas alojadas en estos.

d) Definir la prioridad de rastreo de las URLs, determinando el orden en que los crawlers visitan y procesan cada página web.

9. Los algoritmos para detectar comunidades en una red intentan:

a) Buscar conjuntos donde cada nodo de un mismo conjunto tenga características similares al resto de los nodos del conjunto.

b) Seleccionar aleatoriamente nodos de alto grado y sus vecinos.

c) Buscar subgrafos tal que no incluyan nodos cuya ausencia desconecte al subgrafo.

d) Encontrar grupos donde los nodos pertenecientes a los mismos grupos son cercanos bajo cierta métrica y lejanos con respecto a los nodos de otros grupos.

10. Al "relajar" el concepto de clique en la detección de comunidades se intenta solucionar:

a) La imposibilidad de conectar nodos distantes.

b) La necesidad de datos externos para analizar la red.

c) La uniformidad de los nodos en términos de grado.

d) El solapamiento y la complejidad computacional.

11. El algoritmo Hypertext Induced Topic Selection (HITS) intenta buscar nodos especiales. Estos son conocidos como:

a) Autoridades.

b) Hubs.

c) Centrales.

d) Sensibles.

12. La web actual se enfrenta a problemas como:

a) Presencia de grandes volúmenes de datos estructurados.

b) Presencia de una alta calidad en los datos.

c) Presencia de datos volátiles y distribuidos.

d) Heterogeneidad en los datos.

13. Con respecto a WordNet y su contribución a las ontologías y la representación del conocimiento en los sistemas de RI, puede afirmarse que:

a) En WordNet, una palabra está asociada a un conjunto de sinónimos (synsets), siendo estas palabras intercambiables en un contexto.

b) La integración de WordNet en sistemas de RI permite la expansión de consultas y la mejora de la precisión de los resultados al entender mejor el significado de los términos de búsqueda a través de su contexto semántico y las relaciones entre palabras.

c) WordNet diferencia claramente entre los significados de palabras según su uso en diferentes contextos, lo que permite aplicaciones avanzadas en desambiguación semántica más allá de los sistemas de recomendación y RI.

d) Una limitación de WordNet en la representación del conocimiento es su enfoque en el idioma inglés, lo que plantea desafíos en la aplicación global y la interoperabilidad con sistemas de información multilingües.

14. La integración de Hadoop y MapReduce en la RI trae como ventaja:

a) La eliminación de la necesidad de sistemas de bases de datos.

b) La posibilidad de la extracción de información relevante y la generación de resultados significativos de grandes conjuntos de datos.

c) La garantía de la privacidad absoluta de los datos procesados.

d) La reducción de los costos operativos a cero.

15. En el contexto de la RI en redes sociales se puede afirmar que:

a) El análisis de redes sociales se centra exclusivamente en la cantidad de seguidores que tiene un usuario en particular para determinar su influencia en la red.

b) El análisis de centralidad de intermediación se utiliza para identificar usuarios que son importantes en una red social debido a su posición como "puentes" entre diferentes grupos de usuarios.

c) El análisis de sentimientos se utiliza para determinar la popularidad de una publicación en redes sociales sin tener en cuenta la opinión de los usuarios.

d) El análisis de redes sociales no es útil para comprender la difusión de información en una red social específica, para ello se utiliza la medida de centralidad de vector propio.

16. Una plataforma de comercio electrónico desea mejorar la experiencia del usuario al permitir una navegación más personalizada y contextualizada. Actualmente los usuarios tienen dificultades para encontrar productos relevantes debido a la gran cantidad de opciones disponibles. La empresa está interesada en implementar características de la Web 2.5 y la Web

Semántica para abordar este problema. ¿Qué características podrían ayudar para ofrecer una navegación más personalizada y contextualizada?

a) La implementación de ontologías y metadatos para enriquecer la descripción de productos y mejorar la precisión de las recomendaciones.

b) La optimización de la velocidad de carga del sitio web para mejorar la experiencia del usuario y reducir el abandono del carrito de compra.

c) La integración de redes sociales para permitir la recomendación de productos basada en las preferencias de amigos y contactos.

d) La incorporación de los productos en tendencia en el mercado.

17. En el algoritmo de Indexación Basada en Clasificación Bloqueada (BSBI), ¿cuál es el paso final para crear un índice invertido para la colección completa de los datos?

a) Eliminar los términos duplicados de los índices.

b) Indexar cada bloque de forma independiente.

c) Fusionar los índices invertidos de cada bloque.

d) Dividir la colección de datos en bloques de tamaño fijo.

18. En el contexto del procesamiento de grandes conjuntos de datos, se puede asegurar que:

a) MapReduce es un enfoque para procesar datos de forma secuencial en un solo servidor para evitar problemas de concurrencia.

b) MapReduce divide una tarea en múltiples pasos de map y reduce que se ejecutan de forma secuencial en diferentes servidores para mejorar el rendimiento y la escalabilidad.

c) MapReduce no es adecuado para el procesamiento de datos no estructurados.

d) MapReduce solo puede manejar pequeñas cantidades de datos y no escala bien a grandes conjuntos de datos.

19. El algoritmo de PageRank converge si:

a) La norma de la diferencia entre los vectores es menor a un umbral predefinido.

b) Finaliza la ejecución cuando el número de iteraciones excede un máximo de iteraciones previamente definido.

c) Se define un factor de normalización en la fórmula de la función.

d) El algoritmo no se implementa de forma iterativa.

20. Sobre el algoritmo de PageRank, visto en clase, se puede afirmar que:

a) Evalúa la importancia de un sitio web en función de la calidad y cantidad de enlaces entrantes que recibe de otros sitios web.

b) Solo tiene en cuenta el contenido en un sitio web para determinar su relevancia en los resultados de búsqueda.

c) Asigna una puntuación alta a los sitios web que tienen un gran número de enlaces entrantes sin tener en cuenta la calidad de esos enlaces.

d) Asigna una puntuación baja a los sitios web que contienen muchos enlaces salientes, ya que indica una falta de relevancia.

Trabajo de Control
Sistemas de Recuperación de Información
Fecha: 15 de julio de 2024
Temario #44

Nombre: _____ Grupo: _____

Seleccione en cada caso la(s) respuesta(s) correcta(s). Considere también no marcar ninguna opción.

1. En el diseño de un videojuego de roles (RPG) en el que los personajes tienen atributos como salud, fuerza y velocidad, ¿cuál de las siguientes opciones representa mejor una implementación de la representación del conocimiento orientado a objetos?
   a) Cada personaje se representa como una lista de cadenas de texto que describen sus características físicas y habilidades.
   b) Cada personaje se representa como una función que calcula sus atributos en función de su nivel y experiencia.
   c) Cada personaje se representa como una matriz de números que almacena sus valores de atributos.
   d) Cada personaje se representa como un objeto con propiedades como salud, fuerza y velocidad, y métodos para modificar y consultar estos valores.

2. Referente al Web Crawler puede afirmarse que:
   a) Los hipervínculos encontrados en cada sitio web que no pertenecen al dominio donde fueron encontrados se desechan, puesto que no expande el conjunto de URLs sin visitar.
   b) Las páginas visitadas no se procesan nunca más.
   c) No necesita de un conjunto inicial de URLs para recorrer la Web.
   d) No tiene como objetivo indexar y recopilar información de diferentes sitios web.

3. En un grafo una comunidad es:
   a) Un conjunto de nodos que no comparten ninguna similitud estructural o funcional entre sí.
   b) Un conjunto de nodos aislados.
   c) Un conjunto de nodos altamente conectados que forman un subgrafo completamente independiente del resto de la red.
   d) Un conjunto de nodos que están más densamente interconectados entre sí que con los nodos fuera del conjunto.

4. Al realizar la optimización de contenido para SEO debe considerarse:
   a) Seleccionar las palabras clave al azar.
   b) Utilizar etiquetas de título y meta descripciones únicas y relevantes para cada página.
   c) Incluir palabras clave de manera excesiva en el contenido para mejorar el posicionamiento.
   d) Crear contenido valioso y original que satisfaga las necesidades de los usuarios.

5. La diferencia entre la indexación por tokens y la indexación por conceptos puede definirse como:
   a) La indexación por tokens asigna pesos a los términos basados en su importancia relativa, mientras que la indexación por conceptos utiliza un sistema de etiquetado para asociar términos con características generales.
   b) La indexación por tokens divide los datos en términos individuales, mientras que la indexación por conceptos agrupa los datos en categorías definidas.
   c) La indexación por tokens asigna un valor numérico a cada término de los datos, mientras que la indexación por conceptos utiliza algoritmos de encriptación para proteger la privacidad de los datos.
   d) La indexación por tokens normaliza los datos reduciéndolos a su forma básica, mientras que la indexación por conceptos utiliza un método de ordenación para organizar los términos característicos de los datos.

6. El algoritmo de PageRank converge si:
   a) La norma de la diferencia entre los vectores es menor a un umbral predefinido.
   b) Finaliza la ejecución cuando el número de iteraciones excede un máximo de iteraciones previamente definido.
   c) Se define un factor de normalización en la fórmula de la función.
   d) El algoritmo no se implementa de forma iterativa.

7. Una plataforma de comercio electrónico desea mejorar la experiencia del usuario al permitir una navegación más personalizada y contextualizada. Actualmente los usuarios tienen dificultades para encontrar productos relevantes debido a la gran cantidad de opciones disponibles. La empresa está interesada en implementar características de la Web 2.5 y la Web Semántica para abordar este problema. ¿Qué características podrían ayudar para ofrecer una navegación más personalizada y contextualizada?

a) La implementación de ontologías y metadatos para enriquecer la descripción de productos y mejorar la precisión de las recomendaciones.

b) La optimización de la velocidad de carga del sitio web para mejorar la experiencia del usuario y reducir el abandono del carrito de compra.

c) La integración de redes sociales para permitir la recomendación de productos basada en las preferencias de amigos y contactos.

d) La incorporación de los productos en tendencia en el mercado.

8. No se considera como técnica para detectar comunidades en una red:

a) Analizar la mutualidad de los enlaces.

b) Usar el agrupamiento jerárquico.

c) Utilizar el algoritmo de K-Means.

d) Encontrar cliques de vértices de grado par.

9. En un sistema donde el conocimiento está orientado a objetos se puede asegurar que:

a) Se enfatiza la atención a la información de la cual se extrajo el conocimiento.

b) Los marcos y las bandas son las estructuras utilizadas para la representación del modelo.

c) No es posible definir especializaciones de los objetos de la vida real dentro del sistema.

d) Las funciones de agregación dificultan poder establecer relaciones entre los objetos.

10. En el contexto de la RI en redes sociales se puede afirmar que:

a) El análisis de redes sociales se centra exclusivamente en la cantidad de seguidores que tiene un usuario en particular para determinar su influencia en la red.

b) El análisis de centralidad de intermediación se utiliza para identificar usuarios que son importantes en una red social debido a su posición como "puentes" entre diferentes grupos de usuarios.

c) El análisis de sentimientos se utiliza para determinar la popularidad de una publicación en redes sociales sin tener en cuenta la opinión de los usuarios.

d) El análisis de redes sociales no es útil para comprender la difusión de información en una red social específica, para ello se utiliza la medida de centralidad de vector propio.

11. Un investigador necesita recopilar datos de múltiples sitios web para un estudio académico, pero se enfrenta a varios desafíos al realizar el proceso de Web Scraping de manera ética y legal. ¿Cuál de las siguientes opciones describe mejor uno de los desafíos asociados al proceso de Web Scraping?

a) La necesidad de comprender la estructura del sitio web y su código HTML para extraer los datos correctamente.

b) La disponibilidad limitada de datos en línea que se pueden extraer utilizando técnicas de Web Scraping.

c) La dificultad para encontrar herramientas de Web Scraping gratuitas y fiables.

d) La necesidad de estar montado sobre un scrawler que cumpla con todas las políticas.

12. ¿Qué estrategia utilizan los Web Crawlers para asegurar un rastreo eficiente y respetuoso de los recursos de los sitios web?

a) Ignorar completamente el archivo `robots.txt` de los sitios web.

b) Visitar y rastrear todos los enlaces de una página web simultáneamente.

c) Extraer únicamente contenido multimedia para reducir la carga en los servidores web.

d) Seguir las directrices del archivo `robots.txt` y aplicar un retraso entre las solicitudes.

13. ¿Qué es un "Uniform Resource Locator (URL) Frontier" en el contexto de Web Crawling?

a) Un protocolo que define cómo se deben formatear las URLs para el crawling.

b) Una lista prioritaria de URLs que aún no han sido visitadas por el crawler.

c) Una técnica para filtrar URLs irrelevantes y mejorar la eficiencia del crawling.

d) Una base de datos que almacena URLs únicas identificadas como recursos en la web.

14. El algoritmo Hypertext Induced Topic Selection (HITS) intenta buscar nodos especiales. Estos son conocidos como:

a) Autoridades.

b) Hubs.

c) Centrales.

d) Sensibles.

15. Existen varias estrategias consideradas como no recomendables para mejorar el SEO de un sitio web. Dentro de estas estrategias negativas, se encuentran:

a) Construir enlaces naturales de sitios web con autoridad y relevancia temática.

b) Evitar el uso de las meta etiquetas y descripciones del sitio web para reflejar el contenido de la página.

c) Incluir una cantidad excesiva de palabras clave irrelevantes para intentar manipular los rankings de búsqueda.

d) Utilizar técnicas de encubrimiento para mostrar contenido diferente a lo indexado por los motores de búsqueda.

16. La afirmación que mejor refleja el principio subyacente de PageRank, considerando su importancia en la clasificación de los sitios web, es:
    a) PageRank valora más la cantidad de enlaces entrantes a una página web, independientemente de la calidad o relevancia de estos enlaces.
    b) La efectividad de PageRank se basa exclusivamente en el análisis de las palabras clave contenidas en los enlaces entrantes, sin considerar la estructura de enlace de la Web.
    c) El algoritmo de PageRank considera tanto la cantidad como la calidad de los enlaces entrantes, asignando mayor valor a los enlaces provenientes de sitios web considerados como "importantes".
    d) PageRank opera bajo el supuesto de que los enlaces entrantes y salientes tienen el mismo impacto en la valoración de la relevancia de una página web.

17. En una empresa comercial, se tiene una red donde los nodos corresponden a los empleados y las aristas representan la frecuencia con la que colaboran en las ventas. Se busca mejorar la colaboración entre los empleados para aumentar las ventas totales, por lo que la directiva debe:
    a) Utilizar la centralidad de cercanía para identificar los empleados menos cercanos a otros en la red y asignarles tareas individuales para evitar posibles conflictos y desacuerdos en el proceso de colaboración.
    b) Utilizar la centralidad de cercanía para identificar los empleados más cercanos a otros en la red y promover la colaboración entre ellos, facilitando así la comunicación y el intercambio de conocimientos para mejorar las ventas.
    c) No considerar la centralidad de cercanía de los empleados al diseñar estrategias de colaboración, ya que todos los empleados tienen el mismo acceso a la información y los recursos.
    d) No tener en cuenta la centralidad de cercanía de los empleados al diseñar estrategias de colaboración, ya que esto podría introducir complicaciones adicionales en el proceso de trabajo.

18. En la RI en el contexto de Big Data, se puede asegurar que:
    a) El almacenamiento distribuido centraliza todos los datos en un único servidor para facilitar su gestión y mantenimiento.
    b) El almacenamiento distribuido ofrece ventajas significativas en términos de escalabilidad y rendimiento en comparación con el almacenamiento centralizado.
    c) El almacenamiento distribuido reparte los datos en múltiples servidores para mejorar la disponibilidad y la redundancia del sistema.
    d) El almacenamiento distribuido es una técnica obsoleta.

19. Una buena práctica de SEO para mejorar el posicionamiento de un sitio web en los motores de búsqueda es:
    a) Obtener enlaces de otros sitios web relevantes y de calidad que apunten al sitio.
    b) Copiar contenido directamente de otros sitios web populares para aumentar la cantidad de páginas indexadas.
    c) Utilizar técnicas de encubrimiento para mostrar contenido diferente a lo indexado por los motores de búsqueda.
    d) Llenar el contenido con palabras clave irrelevantes para aumentar la densidad de palabras clave.

20. La integración de Hadoop y MapReduce en la RI trae como ventaja:
    a) La eliminación de la necesidad de sistemas de bases de datos.
    b) La posibilidad de la extracción de información relevante y la generación de resultados significativos de grandes conjuntos de datos.
    c) La garantía de la privacidad absoluta de los datos procesados.
    d) La reducción de los costos operativos a cero.

Trabajo de Control
Sistemas de Recuperación de Información
Fecha: 15 de julio de 2024
Temario #45

Nombre: _____ Grupo: _____

Seleccione en cada caso la(s) respuesta(s) correcta(s). Considere también no marcar ninguna opción.

1. La Web 1.0 se caracteriza por:
   a) Las actualizaciones del contenido de los sitios web se efectúa de forma manual dentro del código HTML
   b) Los sitios web se centran en brindar información en lugar de facilitar la colaboración o participación de los usuarios.
   c) Uso de comunidades virtuales para popularizar los sitios web de noticias.
   d) Los propietarios de los sitios web proporcionan contenido de forma periódica.

2. Dentro del ecosistema de Hadoop, el HDFS se caracteriza por:
   a) El modelo de acceso y de escritura de datos en tiempo real.
   b) La tolerancia a fallos mediante la replicación de datos.
   c) El almacenamiento exclusivo para archivos de texto.
   d) La capacidad ilimitada de almacenamiento.

3. Sobre el algoritmo de PageRank, visto en clase, se puede afirmar que:
   a) Evalúa la importancia de un sitio web en función de la calidad y cantidad de enlaces entrantes que recibe de otros sitios web.
   b) Solo tiene en cuenta el contenido en un sitio web para determinar su relevancia en los resultados de búsqueda.
   c) Asigna una puntuación alta a los sitios web que tienen un gran número de enlaces entrantes sin tener en cuenta la calidad de esos enlaces.
   d) Asigna una puntuación baja a los sitios web que contienen muchos enlaces salientes, ya que indica una falta de relevancia.

4. La web actual se enfrenta a problemas como:
   a) Presencia de grandes volúmenes de datos estructurados.
   b) Presencia de una alta calidad en los datos.
   c) Presencia de datos volátiles y distribuidos.
   d) Heterogeneidad en los datos.

5. En una red de transporte donde cada nodo es una parada de autobús y las aristas representan si existe un carro que pasa por ambos sitios, ¿qué puede mejorar el sistema de transporte?
   a) No considerar la centralidad de intermediación de las estaciones de transporte público al planificar rutas y horarios, ya que no tiene impacto en las conexiones entre las paradas.
   b) Utilizar el grafo inducido de los nodos con mayor valor en la centralidad de grado para aplicar la centralidad de intermediación con el propósito de reforzar las paradas con mayor tráfico.
   c) Utilizar la centralidad de intermediación para identificar las paradas de transferencia clave y establecer nuevas conexiones entre diferentes líneas de transporte público.
   d) No implementar sistemas de información en tiempo real para los usuarios, ya que pueden aumentar la carga de los trabajadores.

6. En un sistema donde el conocimiento está definido a partir de reglas se puede asegurar que:
   a) El nivel de especificidad de las reglas está limitado.
   b) La representación de conocimiento está basada en la lógica proposicional.
   c) El razonamiento se activa en cadena hacia delante.
   d) El orden en que se definen las reglas no altera el razonamiento del sistema.

7. Se tiene un grafo donde cada nodo es un personaje de cierto libro de cuentos y la existencia de las aristas está definida si dos personajes aparecen en el mismo cuento. Se puede asegurar que:
   a) La centralidad de intermediación identifica a los personajes que actúan como conectores entre personajes de cuentos distintos.
   b) La centralidad de grado es útil para identificar los personajes con más conexiones dentro de la red, lo que puede indicar su importancia en el libro.
   c) La centralidad de cercanía ofrece una relación entre la cantidad de vecinos de un nodo con respecto a la longitud máxima de un camino dentro del grafo partiendo del nodo en cuestión.

d) La centralidad de cercanía indica el grado de conexión de cada personaje con el resto de los personajes de los cuentos del libro.

8. En la RI en el contexto de Big Data, se puede asegurar que:
   a) La indexación distribuida divide los datos en múltiples fragmentos que se almacenan en varios nodos para permitir búsquedas paralelas y mejorar la escalabilidad.
   b) La indexación distribuida no ofrece ventajas en términos de rendimiento y escalabilidad en comparación con la indexación centralizada.
   c) La indexación distribuida es una técnica obsoleta.
   d) La indexación distribuida almacena todos los datos en un solo servidor para facilitar su acceso y búsqueda.

9. En un grafo una comunidad es:
   a) Un conjunto de nodos que no comparten ninguna similitud estructural o funcional entre sí.
   b) Un conjunto de nodos aislados.
   c) Un conjunto de nodos altamente conectados que forman un subgrafo completamente independiente del resto de la red.
   d) Un conjunto de nodos que están más densamente interconectados entre sí que con los nodos fuera del conjunto.

10. Considerando las dimensiones y desafíos inherentes a Big Data puede afirmarse, tomando en cuenta las características clave y las implicaciones para su procesamiento y análisis, que:
    a) Big Data se caracteriza principalmente por su pequeño volumen y uniformidad, permitiendo un procesamiento eficiente con mínimas adaptaciones de las herramientas de análisis de datos tradicionales.
    b) Big Data no desafía la capacidad de las herramientas tradicionales de procesamiento de datos para capturar, almacenar, gestionar y analizar efectivamente la información, dada la evolución constante de las capacidades computacionales y algoritmos de optimización.
    c) Aunque Big Data puede incluir datos estructurados, su naturaleza se expande al incorporar grandes cantidades de datos no estructurados y semiestructurados, lo que exige el uso de tecnologías especializadas en almacenamiento y procesamiento como Hadoop y sistemas de bases de datos NoSQL.
    d) Además de su complejidad y diversidad, Big Data introduce desafíos significativos en términos de veracidad y variabilidad al requerir métodos avanzados de limpieza y validación de datos para asegurar la integridad del análisis.

11. Los algoritmos para detectar comunidades en una red intentan:
    a) Buscar conjuntos donde cada nodo de un mismo conjunto tenga características similares al resto de los nodos del conjunto.
    b) Seleccionar aleatoriamente nodos de alto grado y sus vecinos.
    c) Buscar subgrafos tal que no incluyan nodos cuya ausencia desconecte al subgrafo.
    d) Encontrar grupos donde los nodos pertenecientes a los mismos grupos son cercanos bajo cierta métrica y lejanos con respecto a los nodos de otros grupos.

12. La política de ordenación de URLs en los Web Crawlers tiene como aspecto fundamental:
    a) Limitar el acceso de los crawlers a ciertas secciones de un sitio web, evitando el rastreo de URLs consideradas menos importantes o sensibles.
    b) Definir la estructura de la URL de destino, asegurando que estén ordenadas alfabéticamente para facilitar la navegación y la indexación.
    c) Establecer la prioridad de rastreo de las URLs, determinando el orden en que los crawlers visitan y procesan cada página web.
    d) Determinar la forma en que los crawlers asignan un valor de relevancia a cada URL para clasificarlas en el índice de búsqueda.

13. Un investigador necesita recopilar datos de múltiples sitios web para un estudio académico, pero se enfrenta a varios desafíos al realizar el proceso de Web Scraping de manera ética y legal. ¿Cuál de las siguientes opciones describe mejor uno de los desafíos asociados al proceso de Web Scraping?
    a) La necesidad de comprender la estructura del sitio web y su código HTML para extraer los datos correctamente.
    b) La disponibilidad limitada de datos en línea que se pueden extraer utilizando técnicas de Web Scraping.
    c) La dificultad para encontrar herramientas de Web Scraping gratuitas y fiables.
    d) La necesidad de estar montado sobre un scrawler que cumpla con todas las políticas.

14. Un sistema cuenta con la siguiente información:
    - Juan pertenece a la marina.
    - Juan es capellán (sacerdote encargado de una tarea específica fuera de la parroquia).
    - Los infantes de la marina suelen ser bebedores de cerveza.
    - Un capellán no suele ser bebedor de cerveza.
    - Un bebedor de cerveza suele tener sobrepeso.

- Por lo general, un infante de la marina está en buena condición física.

Si se representa la información en una red de herencia se puede concluir que:

a) No se puede asegurar que Juan sea capellán.

b) No se puede asegurar que Juan esté en buena condición física.

c) Juan tiene sobrepeso.

d) Existen dos razonamientos cancelables.

15. La integración de Hadoop y MapReduce en la RI trae como ventaja:

a) La eliminación de la necesidad de sistemas de bases de datos.

b) La posibilidad de la extracción de información relevante y la generación de resultados significativos de grandes conjuntos de datos.

c) La garantía de la privacidad absoluta de los datos procesados.

d) La reducción de los costos operativos a cero.

16. En un sistema donde el conocimiento está orientado a objetos se puede asegurar que:

a) Se enfatiza la atención a la información de la cual se extrajo el conocimiento.

b) Los marcos y las bandas son las estructuras utilizadas para la representación del modelo.

c) No es posible definir especializaciones de los objetos de la vida real dentro del sistema.

d) Las funciones de agregación dificultan poder establecer relaciones entre los objetos.

17. Si se tiene el conjunto de páginas interconectadas $\{A \rightarrow B, C; B \rightarrow C; C \rightarrow A\}$, entonces la página con valor más alto de PageRank es:

a) A y B.

b) A.

c) C.

d) B.

18. En una biblioteca digital se necesita implementar un sistema que permita a los usuarios encontrar libros y artículos científicos de forma rápida y precisa. Los documentos están en diversos formatos, incluyendo PDF, EPUB y HTML. Se requiere seleccionar un algoritmo de indexación adecuado para el sistema, por lo que el programador designado para la implementación debe considerar:

a) La velocidad de indexación y recuperación de datos.

b) La capacidad para manejar documentos en diferentes formatos.

c) La complejidad del algoritmo en términos de implementación y mantenimiento.

d) La capacidad del algoritmo para procesar imágenes incrustadas o referenciadas en los ficheros.

19. Para la RI, el análisis de las redes puede:

a) Ayudar a identificar grupos de interés.

b) Indicar la importancia de una entidad en la transmisión de la información.

c) Ayudar a comprender la conectividad y la accesibilidad entre las entidades.

d) Revelar patrones de influencia dentro de una comunidad.

20. El concepto de *rank sink* en el algoritmo de PageRank representa páginas web:

a) Con distribuciones uniformes de la puntuación de PageRank.

b) Que tienen una puntuación de PageRank más alta que otras debido a la manipulación de enlaces entrantes y salientes.

c) Con una baja calidad de contenido y una cantidad insuficiente de enlaces salientes, lo que las hace menos relevantes en los resultados de búsqueda.

d) Con un alto número de enlaces salientes que no reciben enlaces entrantes, lo que puede afectar negativamente su puntuación de PageRank.

Trabajo de Control
Sistemas de Recuperación de Información
Fecha: 15 de julio de 2024
Temario #46

Nombre: _____ Grupo: _____

Seleccione en cada caso la(s) respuesta(s) correcta(s). Considere también no marcar ninguna opción.

1. En un sistema donde el conocimiento está orientado a objetos se puede asegurar que:
   a) Se enfatiza la atención a la información de la cual se extrajo el conocimiento.
   b) Los marcos y las bandas son las estructuras utilizadas para la representación del modelo.
   c) No es posible definir especializaciones de los objetos de la vida real dentro del sistema.
   d) Las funciones de agregación dificultan poder establecer relaciones entre los objetos.

2. Considerando las prácticas éticas y legales en el Web Scraping, se puede asegurar que:
   a) Web Scraping requiere considerar las políticas de `robots.txt` del sitio web objetivo.
   b) Es importante revisar y respetar los términos de servicio del sitio web, así como las leyes aplicables de protección de datos y derechos de autor, antes de realizar Web Scraping.
   c) Web Scraping sobre datos personales sin consentimiento es generalmente aceptado si los datos se utilizan con fines de investigación.
   d) La extracción de datos mediante Web Scraping siempre es legal, independientemente de las leyes locales sobre derechos de autor y privacidad de datos definidos en los sitios web.

3. ¿Qué es un "Uniform Resource Locator (URL) Frontier" en el contexto de Web Crawling?
   a) Un protocolo que define cómo se deben formatear las URLs para el crawling.
   b) Una lista prioritaria de URLs que aún no han sido visitadas por el crawler.
   c) Una técnica para filtrar URLs irrelevantes y mejorar la eficiencia del crawling.
   d) Una base de datos que almacena URLs únicas identificadas como recursos en la web.

4. El propósito de la política de revisitado en los Web Crawlers es:
   a) Determinar la frecuencia con la que los crawlers deben volver a visitar una URL específica para mantener la información actualizada en el índice de búsqueda.
   b) Establecer reglas sobre el tiempo máximo que los crawlers pueden pasar en un sitio web durante cada visita para evitar sobrecargar los servidores.
   c) Limitar el acceso de los crawlers a ciertos servidores luego de visitar las páginas alojadas en estos.
   d) Definir la prioridad de rastreo de las URLs, determinando el orden en que los crawlers visitan y procesan cada página web.

5. En el contexto de la RI en redes sociales se puede afirmar que:
   a) El análisis de redes sociales se centra exclusivamente en la cantidad de seguidores que tiene un usuario en particular para determinar su influencia en la red.
   b) El análisis de centralidad de intermediación se utiliza para identificar usuarios que son importantes en una red social debido a su posición como "puentes" entre diferentes grupos de usuarios.
   c) El análisis de sentimientos se utiliza para determinar la popularidad de una publicación en redes sociales sin tener en cuenta la opinión de los usuarios.
   d) El análisis de redes sociales no es útil para comprender la difusión de información en una red social específica, para ello se utiliza la medida de centralidad de vector propio.

6. En una empresa comercial, se tiene una red donde los nodos corresponden a los empleados y las aristas representan la frecuencia con la que colaboran en las ventas. Se busca mejorar la colaboración entre los empleados para aumentar las ventas totales, por lo que la directiva debe:
   a) Utilizar la centralidad de cercanía para identificar los empleados menos cercanos a otros en la red y asignarles tareas individuales para evitar posibles conflictos y desacuerdos en el proceso de colaboración.
   b) Utilizar la centralidad de cercanía para identificar los empleados más cercanos a otros en la red y promover la colaboración entre ellos, facilitando así la comunicación y el intercambio de conocimientos para mejorar las ventas.
   c) No considerar la centralidad de cercanía de los empleados al diseñar estrategias de colaboración, ya que todos los empleados tienen el mismo acceso a la información y los recursos.

d) No tener en cuenta la centralidad de cercanía de los empleados al diseñar estrategias de colaboración, ya que esto podría introducir complicaciones adicionales en el proceso de trabajo.

7. El algoritmo de PageRank converge si:
   a) La norma de la diferencia entre los vectores es menor a un umbral predefinido.
   b) Finaliza la ejecución cuando el número de iteraciones excede un máximo de iteraciones previamente definido.
   c) Se define un factor de normalización en la fórmula de la función.
   d) El algoritmo no se implementa de forma iterativa.
8. Los Web Crawlers se enfrentan a desafíos constantes. Dentro de ellos se encuentran:
   a) La dificultad para generar el contenido dinámico en tiempo real.
   b) La modificación del código y la estructura del sitio web.
   c) La incapacidad para interpretar correctamente el lenguaje de programación utilizado en el desarrollo de los sitios web.
   d) La falta de acceso a la base de datos del servidor web para extraer información actualizada.
9. Una buena práctica de SEO para mejorar el posicionamiento de un sitio web en los motores de búsqueda es:
   a) Obtener enlaces de otros sitios web relevantes y de calidad que apunten al sitio.
   b) Copiar contenido directamente de otros sitios web populares para aumentar la cantidad de páginas indexadas.
   c) Utilizar técnicas de encubrimiento para mostrar contenido diferente a lo indexado por los motores de búsqueda.
   d) Llenar el contenido con palabras clave irrelevantes para aumentar la densidad de palabras clave.
10. ¿Qué estrategia utilizan los Web Crawlers para asegurar un rastreo eficiente y respetuoso de los recursos de los sitios web?
    a) Ignorar completamente el archivo `robots.txt` de los sitios web.
    b) Visitar y rastrear todos los enlaces de una página web simultáneamente.
    c) Extraer únicamente contenido multimedia para reducir la carga en los servidores web.
    d) Seguir las directrices del archivo `robots.txt` y aplicar un retraso entre las solicitudes.
11. En el contexto de la RI en redes, se puede afirmar que:
    a) La detección de comunidades no es relevante para la RI en redes, ya que se centra únicamente en la estructura de la red sin considerar el contenido.
    b) La detección de comunidades en una red siempre produce resultados objetivos y consistentes independientemente del algoritmo utilizado.
    c) La detección de comunidades ayuda a identificar grupos de nodos altamente conectados entre sí, lo que puede ser útil para comprender la estructura y el contenido de la red.
    d) La detección de comunidades solo se aplica a redes pequeñas y simples, no a redes grandes y complejas.
12. La premisa básica del algoritmo de PageRank para clasificar páginas web en los resultados de búsqueda es:
    a) La antigüedad de la página web es el principal factor para determinar su clasificación.
    b) El contenido y la relevancia de las palabras clave en la página web determinan su posición.
    c) Los enlaces entrantes a una página web desde otras páginas contribuyen a su importancia y clasificación.
    d) La cantidad de visitas que recibe una página web determina su clasificación.
13. En la RI en el contexto de Big Data, se puede asegurar que:
    a) El almacenamiento distribuido centraliza todos los datos en un único servidor para facilitar su gestión y mantenimiento.
    b) El almacenamiento distribuido ofrece ventajas significativas en términos de escalabilidad y rendimiento en comparación con el almacenamiento centralizado.
    c) El almacenamiento distribuido reparte los datos en múltiples servidores para mejorar la disponibilidad y la redundancia del sistema.
    d) El almacenamiento distribuido es una técnica obsoleta.
14. En el contexto del procesamiento de grandes conjuntos de datos, se puede asegurar que:
    a) MapReduce es un enfoque para procesar datos de forma secuencial en un solo servidor para evitar problemas de concurrencia.
    b) MapReduce divide una tarea en múltiples pasos de map y reduce que se ejecutan de forma secuencial en diferentes servidores para mejorar el rendimiento y la escalabilidad.
    c) MapReduce no es adecuado para el procesamiento de datos no estructurados.
    d) MapReduce solo puede manejar pequeñas cantidades de datos y no escala bien a grandes conjuntos de datos.

15. Considerando las dimensiones y desafíos inherentes a Big Data puede afirmarse, tomando en cuenta las características clave y las implicaciones para su procesamiento y análisis, que:
    a) Big Data se caracteriza principalmente por su pequeño volumen y uniformidad, permitiendo un procesamiento eficiente con mínimas adaptaciones de las herramientas de análisis de datos tradicionales.
    b) Big Data no desafía la capacidad de las herramientas tradicionales de procesamiento de datos para capturar, almacenar, gestionar y analizar efectivamente la información, dada la evolución constante de las capacidades computacionales y algoritmos de optimización.
    c) Aunque Big Data puede incluir datos estructurados, su naturaleza se expande al incorporar grandes cantidades de datos no estructurados y semiestructurados, lo que exige el uso de tecnologías especializadas en almacenamiento y procesamiento como Hadoop y sistemas de bases de datos NoSQL.
    d) Además de su complejidad y diversidad, Big Data introduce desafíos significativos en términos de veracidad y variabilidad al requerir métodos avanzados de limpieza y validación de datos para asegurar la integridad del análisis.

16. En una biblioteca digital se necesita implementar un sistema que permita a los usuarios encontrar libros y artículos científicos de forma rápida y precisa. Los documentos están en diversos formatos, incluyendo PDF, EPUB y HTML. Se requiere seleccionar un algoritmo de indexación adecuado para el sistema, por lo que el programador designado para la implementación debe considerar:
    a) La velocidad de indexación y recuperación de datos.
    b) La capacidad para manejar documentos en diferentes formatos.
    c) La complejidad del algoritmo en términos de implementación y mantenimiento.
    d) La capacidad del algoritmo para procesar imágenes incrustadas o referenciadas en los ficheros.

17. Implementar índices invertidos en un SRI asegura:
    a) Acelerar el proceso de búsqueda al permitir búsquedas directas por contenido en lugar de por título.
    b) Reducir la cantidad de espacio de almacenamiento necesario al comprimir los datos de los documentos.
    c) Facilitar la búsqueda sobre los datos que contienen términos específicos al mantener una lista de datos para cada término único.
    d) Incrementar la seguridad de los datos almacenados al dificultar el acceso directo a la información sin el índice correcto.

18. En un grafo una comunidad es:
    a) Un conjunto de nodos que no comparten ninguna similitud estructural o funcional entre sí.
    b) Un conjunto de nodos aislados.
    c) Un conjunto de nodos altamente conectados que forman un subgrafo completamente independiente del resto de la red.
    d) Un conjunto de nodos que están más densamente interconectados entre sí que con los nodos fuera del conjunto.

19. En un SRI la indexación:
    a) Consiste en asociar un identificador único a cada dato almacenado en el sistema.
    b) Mejora la experiencia del usuario.
    c) Permite la organización y la categorización de la información.
    d) Facilita la RI relevante.

20. El concepto de *rank sink* en el algoritmo de PageRank representa páginas web:
    a) Con distribuciones uniformes de la puntuación de PageRank.
    b) Que tienen una puntuación de PageRank más alta que otras debido a la manipulación de enlaces entrantes y salientes.
    c) Con una baja calidad de contenido y una cantidad insuficiente de enlaces salientes, lo que las hace menos relevantes en los resultados de búsqueda.
    d) Con un alto número de enlaces salientes que no reciben enlaces entrantes, lo que puede afectar negativamente su puntuación de PageRank.

Trabajo de Control
Sistemas de Recuperación de Información
Fecha: 15 de julio de 2024
Temario #47

Nombre: _____ Grupo: _____

Seleccione en cada caso la(s) respuesta(s) correcta(s). Considere también no marcar ninguna opción.

1. Si se tiene el conjunto de páginas interconectadas $\{A \to B, C; B \to C; C \to A\}$, entonces la página con valor más alto de PageRank es:
   a) A y B.
   b) A.
   c) C.
   d) B.

2. Los Web Crawlers se enfrentan a desafíos constantes. Dentro de ellos se encuentran:
   a) La dificultad para generar el contenido dinámico en tiempo real.
   b) La modificación del código y la estructura del sitio web.
   c) La incapacidad para interpretar correctamente el lenguaje de programación utilizado en el desarrollo de los sitios web.
   d) La falta de acceso a la base de datos del servidor web para extraer información actualizada.

3. La integración de Hadoop y MapReduce en la RI trae como ventaja:
   a) La eliminación de la necesidad de sistemas de bases de datos.
   b) La posibilidad de la extracción de información relevante y la generación de resultados significativos de grandes conjuntos de datos.
   c) La garantía de la privacidad absoluta de los datos procesados.
   d) La reducción de los costos operativos a cero.

4. En la RI en el contexto de Big Data, se puede asegurar que:
   a) La indexación distribuida divide los datos en múltiples fragmentos que se almacenan en varios nodos para permitir búsquedas paralelas y mejorar la escalabilidad.
   b) La indexación distribuida no ofrece ventajas en términos de rendimiento y escalabilidad en comparación con la indexación centralizada.
   c) La indexación distribuida es una técnica obsoleta.
   d) La indexación distribuida almacena todos los datos en un solo servidor para facilitar su acceso y búsqueda.

5. El propósito de la política de revisitado en los Web Crawlers es:
   a) Determinar la frecuencia con la que los crawlers deben volver a visitar una URL específica para mantener la información actualizada en el índice de búsqueda.
   b) Establecer reglas sobre el tiempo máximo que los crawlers pueden pasar en un sitio web durante cada visita para evitar sobrecargar los servidores.
   c) Limitar el acceso de los crawlers a ciertos servidores luego de visitar las páginas alojadas en estos.
   d) Definir la prioridad de rastreo de las URLs, determinando el orden en que los crawlers visitan y procesan cada página web.

6. Implementar índices invertidos en un SRI asegura:
   a) Acelerar el proceso de búsqueda al permitir búsquedas directas por contenido en lugar de por título.
   b) Reducir la cantidad de espacio de almacenamiento necesario al comprimir los datos de los documentos.
   c) Facilitar la búsqueda sobre los datos que contienen términos específicos al mantener una lista de datos para cada término único.
   d) Incrementar la seguridad de los datos almacenados al dificultar el acceso directo a la información sin el índice correcto.

7. ¿Qué es un "Uniform Resource Locator (URL) Frontier" en el contexto de Web Crawling?
   a) Un protocolo que define cómo se deben formatear las URLs para el crawling.
   b) Una lista prioritaria de URLs que aún no han sido visitadas por el crawler.
   c) Una técnica para filtrar URLs irrelevantes y mejorar la eficiencia del crawling.
   d) Una base de datos que almacena URLs únicas identificadas como recursos en la web.

8. En un SRI la indexación:
   a) Consiste en asociar un identificador único a cada dato almacenado en el sistema.
   b) Mejora la experiencia del usuario.
   c) Permite la organización y la categorización de la información.

d) Facilita la RI relevante.

9. No se considera como técnica para detectar comunidades en una red:
   a) Analizar la mutualidad de los enlaces.
   b) Usar el agrupamiento jerárquico.
   c) Utilizar el algoritmo de K-Means.
   d) Encontrar cliques de vértices de grado par.

10. La Web 3 se conoce como:
   a) Internet de las cosas.
   b) Web de solo lectura.
   c) Web semántica.
   d) Web de escritura-lectura.

11. En el contexto de la RI en redes, se puede afirmar que:
   a) La detección de comunidades no es relevante para la RI en redes, ya que se centra únicamente en la estructura de la red sin considerar el contenido.
   b) La detección de comunidades en una red siempre produce resultados objetivos y consistentes independientemente del algoritmo utilizado.
   c) La detección de comunidades ayuda a identificar grupos de nodos altamente conectados entre sí, lo que puede ser útil para comprender la estructura y el contenido de la red.
   d) La detección de comunidades solo se aplica a redes pequeñas y simples, no a redes grandes y complejas.

12. Se puede afirmar que:
   a) El término "Big Data" se refiere exclusivamente al volumen de datos que una organización maneja, sin tener en cuenta la velocidad, la variedad y la veracidad de los datos.
   b) Los SRI pueden beneficiarse de MapReduce para mejorar la RI relevante.
   c) MapReduce es un modelo de procesamiento distribuido utilizado para trabajar con grandes volúmenes de datos.
   d) Uno de los desafíos en el procesamiento de Big Data es la capacidad de gestionar y analizar datos provenientes de diversas fuentes y en diferentes formatos de manera eficiente.

13. En la RI en el contexto de Big Data, se puede asegurar que:
   a) El almacenamiento distribuido centraliza todos los datos en un único servidor para facilitar su gestión y mantenimiento.
   b) El almacenamiento distribuido ofrece ventajas significativas en términos de escalabilidad y rendimiento en comparación con el almacenamiento centralizado.
   c) El almacenamiento distribuido reparte los datos en múltiples servidores para mejorar la disponibilidad y la redundancia del sistema.
   d) El almacenamiento distribuido es una técnica obsoleta.

14. Dentro del análisis de redes, la centralidad de grado mide la importancia de un nodo basándose en:
   a) El grado del nodo.
   b) La cantidad de veces que aparece el nodo en el camino mínimo entre cualquier par de nodos.
   c) La cantidad de vecinos del nodo.
   d) El número de aristas que posee el nodo.

15. Un investigador necesita recopilar datos de múltiples sitios web para un estudio académico, pero se enfrenta a varios desafíos al realizar el proceso de Web Scraping de manera ética y legal. ¿Cuál de las siguientes opciones describe mejor uno de los desafíos asociados al proceso de Web Scraping?
   a) La necesidad de comprender la estructura del sitio web y su código HTML para extraer los datos correctamente.
   b) La disponibilidad limitada de datos en línea que se pueden extraer utilizando técnicas de Web Scraping.
   c) La dificultad para encontrar herramientas de Web Scraping gratuitas y fiables.
   d) La necesidad de estar montado sobre un scrawler que cumpla con todas las políticas.

16. La premisa básica del algoritmo de PageRank para clasificar páginas web en los resultados de búsqueda es:
   a) La antigüedad de la página web es el principal factor para determinar su clasificación.
   b) El contenido y la relevancia de las palabras clave en la página web determinan su posición.
   c) Los enlaces entrantes a una página web desde otras páginas contribuyen a su importancia y clasificación.
   d) La cantidad de visitas que recibe una página web determina su clasificación.

17. La afirmación que mejor refleja el principio subyacente de PageRank, considerando su importancia en la clasificación de los sitios web, es:
   a) PageRank valora más la cantidad de enlaces entrantes a una página web, independientemente de la calidad o relevancia de estos enlaces.
   b) La efectividad de PageRank se basa exclusivamente en el análisis de las palabras clave contenidas en los enlaces entrantes, sin considerar la estructura de enlace de la Web.

c) El algoritmo de PageRank considera tanto la cantidad como la calidad de los enlaces entrantes, asignando mayor valor a los enlaces provenientes de sitios web considerados como "importantes".

d) PageRank opera bajo el supuesto de que los enlaces entrantes y salientes tienen el mismo impacto en la valoración de la relevancia de una página web.

18. La computación evolutiva:

a) No es aplicable en la RI debido a la complejidad de los algoritmos evolutivos.

b) Solo puede manejar conjuntos de datos pequeños y no es escalable a grandes volúmenes de datos.

c) Utiliza algoritmos para buscar soluciones óptimas en grandes espacios de búsqueda, lo que la hace adecuada para problemas de optimización en la RI.

d) Es útil solo para problemas de clasificación de documentos y no para otras tareas de RI en general.

19. Existen varias estrategias consideradas como no recomendables para mejorar el SEO de un sitio web. Dentro de estas estrategias negativas, se encuentran:

a) Construir enlaces naturales de sitios web con autoridad y relevancia temática.

b) Evitar el uso de las meta etiquetas y descripciones del sitio web para reflejar el contenido de la página.

c) Incluir una cantidad excesiva de palabras clave irrelevantes para intentar manipular los rankings de búsqueda.

d) Utilizar técnicas de encubrimiento para mostrar contenido diferente a lo indexado por los motores de búsqueda.

20. La centralidad de intermediación de un nodo indica:

a) La cantidad de nodos vecinos directos.

b) La resistencia del nodo a fallos.

c) El número total de conexiones entrantes y salientes.

d) La frecuencia con la que un nodo actúa como puente en el camino más corto entre otros dos nodos.

Trabajo de Control
Sistemas de Recuperación de Información
Fecha: 15 de julio de 2024
Temario #48

Nombre: _____ Grupo: _____

Seleccione en cada caso la(s) respuesta(s) correcta(s). Considere también no marcar ninguna opción.

1. Considerando las dimensiones y desafíos inherentes a Big Data puede afirmarse, tomando en cuenta las características clave y las implicaciones para su procesamiento y análisis, que:
   a) Big Data se caracteriza principalmente por su pequeño volumen y uniformidad, permitiendo un procesamiento eficiente con mínimas adaptaciones de las herramientas de análisis de datos tradicionales.
   b) Big Data no desafía la capacidad de las herramientas tradicionales de procesamiento de datos para capturar, almacenar, gestionar y analizar efectivamente la información, dada la evolución constante de las capacidades computacionales y algoritmos de optimización.
   c) Aunque Big Data puede incluir datos estructurados, su naturaleza se expande al incorporar grandes cantidades de datos no estructurados y semiestructurados, lo que exige el uso de tecnologías especializadas en almacenamiento y procesamiento como Hadoop y sistemas de bases de datos NoSQL.
   d) Además de su complejidad y diversidad, Big Data introduce desafíos significativos en términos de veracidad y variabilidad al requerir métodos avanzados de limpieza y validación de datos para asegurar la integridad del análisis.

2. En el algoritmo de Indexación Basada en Clasificación Bloqueada (BSBI), ¿cuál es el paso final para crear un índice invertido para la colección completa de los datos?
   a) Eliminar los términos duplicados de los índices.
   b) Indexar cada bloque de forma independiente.
   c) Fusionar los índices invertidos de cada bloque.
   d) Dividir la colección de datos en bloques de tamaño fijo.

3. Al realizar la optimización de contenido para SEO debe considerarse:
   a) Seleccionar las palabras clave al azar.
   b) Utilizar etiquetas de título y meta descripciones únicas y relevantes para cada página.
   c) Incluir palabras clave de manera excesiva en el contenido para mejorar el posicionamiento.
   d) Crear contenido valioso y original que satisfaga las necesidades de los usuarios.

4. La Web 2.0 se caracteriza por:
   a) La existencia de sitios web dinámicos e interactivos que permiten a los usuarios participar, comentar e interactuar tanto con los creadores de contenido como con otros usuarios.
   b) El uso de distintas tecnologías para crear experiencias web más interactivas y con mayor capacidad de respuesta.
   c) Las actualizaciones del contenido de los sitios web se efectúa de forma manual dentro del código HTML
   d) Las plataformas ofrecen experiencias a la medida, permitiendo a los usuarios personalizar sus perfiles, recibir recomendaciones ajustadas al contenido y participar en filtrado colaborativo.

5. Si se tiene el conjunto de páginas interconectadas $\{A \to B, C; B \to C; C \to A\}$, entonces la página con valor más alto de PageRank es:
   a) A y B.
   b) A.
   c) C.
   d) B.

6. El uso de Hadoop y MapReduce en el contexto de la RI tiene como objetivo:
   a) Automatizar el mantenimiento de sistemas de bases de datos.
   b) Facilitar el análisis en tiempo real de datos de redes sociales.
   c) Mejorar la eficiencia energética en centros de datos.
   d) Procesar y analizar grandes conjuntos de datos para la RI.

7. En un SRI la indexación:
   a) Consiste en asociar un identificador único a cada dato almacenado en el sistema.
   b) Mejora la experiencia del usuario.
   c) Permite la organización y la categorización de la información.
   d) Facilita la RI relevante.

8. En un grafo una comunidad es:
   a) Un conjunto de nodos que no comparten ninguna similitud estructural o funcional entre sí.

b) Un conjunto de nodos aislados.

c) Un conjunto de nodos altamente conectados que forman un subgrafo completamente independiente del resto de la red.

d) Un conjunto de nodos que están más densamente interconectados entre sí que con los nodos fuera del conjunto.

9. En un sistema donde el conocimiento está orientado a objetos se puede asegurar que:

a) Se enfatiza la atención a la información de la cual se extrajo el conocimiento.

b) Los marcos y las bandas son las estructuras utilizadas para la representación del modelo.

c) No es posible definir especializaciones de los objetos de la vida real dentro del sistema.

d) Las funciones de agregación dificultan poder establecer relaciones entre los objetos.

10. En el contexto de la RI en redes sociales se puede afirmar que:

a) El análisis de redes sociales se centra exclusivamente en la cantidad de seguidores que tiene un usuario en particular para determinar su influencia en la red.

b) El análisis de centralidad de intermediación se utiliza para identificar usuarios que son importantes en una red social debido a su posición como "puentes" entre diferentes grupos de usuarios.

c) El análisis de sentimientos se utiliza para determinar la popularidad de una publicación en redes sociales sin tener en cuenta la opinión de los usuarios.

d) El análisis de redes sociales no es útil para comprender la difusión de información en una red social específica, para ello se utiliza la medida de centralidad de vector propio.

11. El propósito de la política de revisitado en los Web Crawlers es:

a) Determinar la frecuencia con la que los crawlers deben volver a visitar una URL específica para mantener la información actualizada en el índice de búsqueda.

b) Establecer reglas sobre el tiempo máximo que los crawlers pueden pasar en un sitio web durante cada visita para evitar sobrecargar los servidores.

c) Limitar el acceso de los crawlers a ciertos servidores luego de visitar las páginas alojadas en estos.

d) Definir la prioridad de rastreo de las URLs, determinando el orden en que los crawlers visitan y procesan cada página web.

12. Sobre el algoritmo de PageRank, visto en clase, se puede afirmar que:

a) Evalúa la importancia de un sitio web en función de la calidad y cantidad de enlaces entrantes que recibe de otros sitios web.

b) Solo tiene en cuenta el contenido en un sitio web para determinar su relevancia en los resultados de búsqueda.

c) Asigna una puntuación alta a los sitios web que tienen un gran número de enlaces entrantes sin tener en cuenta la calidad de esos enlaces.

d) Asigna una puntuación baja a los sitios web que contienen muchos enlaces salientes, ya que indica una falta de relevancia.

13. Un SRI es capaz de:

a) Crear los índices asociados a los datos sin tener que analizar cada dato.

b) Generar índices invertidos de manera óptima sin considerar el contexto.

c) Reducir el tiempo de indexación de los datos si utiliza servidores distribuidos dentro de la red para que cada uno ejecute la indexación del mismo conjunto de datos.

d) No necesitar de ningún almacenamiento externo para alojar los índices de los datos.

14. El componente responsable de la gestión de recursos y planificación de tareas en Hadoop es:

a) MapReduce.

b) Hadoop Common.

c) HDFS.

d) YARN.

15. El concepto de *rank sink* en el algoritmo de PageRank representa páginas web:

a) Con distribuciones uniformes de la puntuación de PageRank.

b) Que tienen una puntuación de PageRank más alta que otras debido a la manipulación de enlaces entrantes y salientes.

c) Con una baja calidad de contenido y una cantidad insuficiente de enlaces salientes, lo que las hace menos relevantes en los resultados de búsqueda.

d) Con un alto número de enlaces salientes que no reciben enlaces entrantes, lo que puede afectar negativamente su puntuación de PageRank.

16. En el contexto de la RI en redes, se puede afirmar que:

a) La detección de comunidades no es relevante para la RI en redes, ya que se centra únicamente en la estructura de la red sin considerar el contenido.

b) La detección de comunidades en una red siempre produce resultados objetivos y consistentes independientemente del algoritmo utilizado.

c) La detección de comunidades ayuda a identificar grupos de nodos altamente conectados entre sí, lo que puede ser útil para comprender la estructura y el contenido de la red.

d) La detección de comunidades solo se aplica a redes pequeñas y simples, no a redes grandes y complejas.

17. Una buena práctica de SEO para mejorar el posicionamiento de un sitio web en los motores de búsqueda es:

a) Obtener enlaces de otros sitios web relevantes y de calidad que apunten al sitio.

b) Copiar contenido directamente de otros sitios web populares para aumentar la cantidad de páginas indexadas.

c) Utilizar técnicas de encubrimiento para mostrar contenido diferente a lo indexado por los motores de búsqueda.

d) Llenar el contenido con palabras clave irrelevantes para aumentar la densidad de palabras clave.

18. Para transformar el contenido no estructurado de las páginas web en datos estructurados, el Web Scraping:

a) Extrae información basada en patrones de HTML/CSS.

b) Analiza los protocolos de red.

c) Interpreta el código JavaScript en tiempo real.

d) Convierte de forma automática imágenes a texto.

19. Sobre WordNet y su aplicación en el procesamiento del lenguaje natural, se puede afirmar que:

a) El diseño de WordNet facilita su integración en aplicaciones multilingües de NLP, aunque su desarrollo original se centró en el inglés.

b) Aunque WordNet es una herramienta valiosa en el NLP, su estructura no incluye información sobre la frecuencia de uso de las palabras en el lenguaje natural.

c) Los synsets facilitan la identificación de relaciones semánticas entre palabras, como la hiperonimia y la meronimia, enriqueciendo tareas de NLP.

d) WordNet proporciona una base para la desambiguación semántica al agrupar palabras con significados similares en synsets.

20. La Web 3 se conoce como:

a) Internet de las cosas.

b) Web de solo lectura.

c) Web semántica.

d) Web de escritura-lectura.

Trabajo de Control
Sistemas de Recuperación de Información
Fecha: 15 de julio de 2024
Temario #49

Nombre: _____ Grupo: _____

Seleccione en cada caso la(s) respuesta(s) correcta(s). Considere también no marcar ninguna opción.

1. La web actual se enfrenta a problemas como:
   a) Presencia de grandes volúmenes de datos estructurados.
   b) Presencia de una alta calidad en los datos.
   c) Presencia de datos volátiles y distribuidos.
   d) Heterogeneidad en los datos.

2. En un grafo una comunidad es:
   a) Un conjunto de nodos que no comparten ninguna similitud estructural o funcional entre sí.
   b) Un conjunto de nodos aislados.
   c) Un conjunto de nodos altamente conectados que forman un subgrafo completamente independiente del resto de la red.
   d) Un conjunto de nodos que están más densamente interconectados entre sí que con los nodos fuera del conjunto.

3. La centralidad de intermediación de un nodo indica:
   a) La cantidad de nodos vecinos directos.
   b) La resistencia del nodo a fallos.
   c) El número total de conexiones entrantes y salientes.
   d) La frecuencia con la que un nodo actúa como puente en el camino más corto entre otros dos nodos.

4. El algoritmo de PageRank puede describirse como un procedimiento utilizado para:
   a) Determinar la velocidad de carga de una página web en un navegador.
   b) Calcular la relevancia de una página web en función de la cantidad y calidad de los enlaces que apuntan hacia ella.
   c) Prevenir el spam y el contenido no deseado en las páginas web.
   d) Clasificar las páginas web en función de su edad y autoridad.

5. El algoritmo de PageRank converge si:
   a) La norma de la diferencia entre los vectores es menor a un umbral predefinido.
   b) Finaliza la ejecución cuando el número de iteraciones excede un máximo de iteraciones previamente definido.
   c) Se define un factor de normalización en la fórmula de la función.
   d) El algoritmo no se implementa de forma iterativa.

6. En un SRI la indexación:
   a) Consiste en asociar un identificador único a cada dato almacenado en el sistema.
   b) Mejora la experiencia del usuario.
   c) Permite la organización y la categorización de la información.
   d) Facilita la RI relevante.

7. En la RI en el contexto de Big Data, se puede asegurar que:
   a) La indexación distribuida divide los datos en múltiples fragmentos que se almacenan en varios nodos para permitir búsquedas paralelas y mejorar la escalabilidad.
   b) La indexación distribuida no ofrece ventajas en términos de rendimiento y escalabilidad en comparación con la indexación centralizada.
   c) La indexación distribuida es una técnica obsoleta.
   d) La indexación distribuida almacena todos los datos en un solo servidor para facilitar su acceso y búsqueda.

8. En un sistema de control de tráfico urbano basado en reglas, ¿cuál de las siguientes reglas sería más efectiva para manejar situaciones de congestión en una intersección?
   a) Si hay pocos vehículos en la intersección, reducir el tiempo de los semáforos en verde.
   b) Si hay muchos vehículos en la intersección, aumentar el tiempo de los semáforos en verde.
   c) Si hay un vehículo de emergencia en la intersección, detener todos los demás vehículos.
   d) Si hay muchos peatones cruzando la intersección, reducir el tiempo de los semáforos en rojo.

9. ¿Qué algoritmos permiten obtener información de una red?
   a) Índices de centralidad.
   b) Detección de comunidades.
   c) Hypertext Induced Topic Selection (HITS).
   d) PageRank.

10. Un SRI es capaz de:
    a) Crear los índices asociados a los datos sin tener que analizar cada dato.
    b) Generar índices invertidos de manera óptima sin considerar el contexto.
    c) Reducir el tiempo de indexación de los datos si utiliza servidores distribuidos dentro de la red para que cada uno ejecute la indexación del mismo conjunto de datos.
    d) No necesitar de ningún almacenamiento externo para alojar los índices de los datos.

11. La Web 2.0 se caracteriza por:
    a) La existencia de sitios web dinámicos e interactivos que permiten a los usuarios participar, comentar e interactuar tanto con los creadores de contenido como con otros usuarios.
    b) El uso de distintas tecnologías para crear experiencias web más interactivas y con mayor capacidad de respuesta.
    c) Las actualizaciones del contenido de los sitios web se efectúa de forma manual dentro del código HTML
    d) Las plataformas ofrecen experiencias a la medida, permitiendo a los usuarios personalizar sus perfiles, recibir recomendaciones ajustadas al contenido y participar en filtrado colaborativo.

12. En la RI en el contexto de Big Data, se puede asegurar que:
    a) El almacenamiento distribuido centraliza todos los datos en un único servidor para facilitar su gestión y mantenimiento.
    b) El almacenamiento distribuido ofrece ventajas significativas en términos de escalabilidad y rendimiento en comparación con el almacenamiento centralizado.
    c) El almacenamiento distribuido reparte los datos en múltiples servidores para mejorar la disponibilidad y la redundancia del sistema.
    d) El almacenamiento distribuido es una técnica obsoleta.

13. Para transformar el contenido no estructurado de las páginas web en datos estructurados, el Web Scraping:
    a) Extrae información basada en patrones de HTML/CSS.
    b) Analiza los protocolos de red.
    c) Interpreta el código JavaScript en tiempo real.
    d) Convierte de forma automática imágenes a texto.

14. La integración de Hadoop y MapReduce en la RI trae como ventaja:
    a) La eliminación de la necesidad de sistemas de bases de datos.
    b) La posibilidad de la extracción de información relevante y la generación de resultados significativos de grandes conjuntos de datos.
    c) La garantía de la privacidad absoluta de los datos procesados.
    d) La reducción de los costos operativos a cero.

15. Referente al Web Crawler puede afirmarse que:
    a) Los hipervínculos encontrados en cada sitio web que no pertenecen al dominio donde fueron encontrados se desechan, puesto que no expande el conjunto de URLs sin visitar.
    b) Las páginas visitadas no se procesan nunca más.
    c) No necesita de un conjunto inicial de URLs para recorrer la Web.
    d) No tiene como objetivo indexar y recopilar información de diferentes sitios web.

16. El algoritmo Hypertext Induced Topic Selection (HITS) intenta buscar nodos especiales. Estos son conocidos como:
    a) Autoridades.
    b) Hubs.
    c) Centrales.
    d) Sensibles.

17. Se tiene una red de ingredientes donde cada uno representa un nodo y las aristas simbolizan que los ingredientes forman parte de una misma receta. Se busca mejorar la experiencia culinaria mediante la elaboración de combinaciones de ingredientes más interesantes y creativas, para lo cual se debe:
    a) No considerar la centralidad de grado de los ingredientes, ya que todas las combinaciones de ingredientes son igualmente válidas.
    b) Utilizar la centralidad de grado para identificar los ingredientes menos conectados en la red y tomarlos en cuenta para su inclusión en las combinaciones.
    c) No tener en cuenta la centralidad de grado de los ingredientes para no darle mayor importancia a los ingredientes más comunes.
    d) Utilizar la centralidad de grado para identificar los ingredientes más populares en la red y crear combinaciones que incluyan una variedad de ingredientes menos comunes.

18. En el contexto de la representación del conocimiento basada en herencia, ¿qué caracteriza a la herencia cancelable?
    a) La herencia es el resultado del razonamiento no transitivo.
    b) Las conclusiones no están determinadas y dependen del nodo de interés.

c) Las propiedades heredadas siempre se mantienen y no pueden anularse.

d) No existe ambigüedad en las conclusiones obtenidas.

19. Con respecto a WordNet y su contribución a las ontologías y la representación del conocimiento en los sistemas de RI, puede afirmarse que:

a) En WordNet, una palabra está asociada a un conjunto de sinónimos (synsets), siendo estas palabras intercambiables en un contexto.

b) La integración de WordNet en sistemas de RI permite la expansión de consultas y la mejora de la precisión de los resultados al entender mejor el significado de los términos de búsqueda a través de su contexto semántico y las relaciones entre palabras.

c) WordNet diferencia claramente entre los significados de palabras según su uso en diferentes contextos, lo que permite aplicaciones avanzadas en desambiguación semántica más allá de los sistemas de recomendación y RI.

d) Una limitación de WordNet en la representación del conocimiento es su enfoque en el idioma inglés, lo que plantea desafíos en la aplicación global y la interoperabilidad con sistemas de información multilingües.

20. La transición de la Web 1.0 a la Web 2.0 se caracterizó principalmente por:

a) El aumento en la velocidad de conexión a internet, que permitió una mejor calidad de las páginas web.

b) La disminución de la importancia de los motores de búsqueda en la navegación web.

c) El cambio de páginas web estáticas a dinámicas, permitiendo la interacción del usuario y la generación de contenido.

d) La reducción en el uso de HTML y CSS en el desarrollo de sitios web.

Trabajo de Control
Sistemas de Recuperación de Información
Fecha: 15 de julio de 2024
Temario #50

Nombre: _____ Grupo: _____

Seleccione en cada caso la(s) respuesta(s) correcta(s). Considere también no marcar ninguna opción.

1. En un sistema donde el conocimiento está definido a partir de reglas se puede asegurar que:
   a) El nivel de especificidad de las reglas está limitado.
   b) La representación de conocimiento está basada en la lógica proposicional.
   c) El razonamiento se activa en cadena hacia delante.
   d) El orden en que se definen las reglas no altera el razonamiento del sistema.

2. Considerando las prácticas éticas y legales en el Web Scraping, se puede asegurar que:
   a) Web Scraping requiere considerar las políticas de `robots.txt` del sitio web objetivo.
   b) Es importante revisar y respetar los términos de servicio del sitio web, así como las leyes aplicables de protección de datos y derechos de autor, antes de realizar Web Scraping.
   c) Web Scraping sobre datos personales sin consentimiento es generalmente aceptado si los datos se utilizan con fines de investigación.
   d) La extracción de datos mediante Web Scraping siempre es legal, independientemente de las leyes locales sobre derechos de autor y privacidad de datos definidos en los sitios web.

3. Analizar una red permite:
   a) Detectar posibles tendencias antes de que se conviertan en tendencia.
   b) Obtener predicciones exactas de eventos futuros en mercados financieros.
   c) Evaluar la calidad del contenido de un sitio web.
   d) Encontrar nodos "sensibles" o críticos para la red.

4. La Web 3 se conoce como:
   a) Internet de las cosas.
   b) Web de solo lectura.
   c) Web semántica.
   d) Web de escritura-lectura.

5. Para contribuir positivamente al posicionamiento orgánico de un sitio web en los motores de búsqueda se puede:
   a) Crear contenido relevante y de alta calidad que satisfaga las necesidades de información de los usuarios.
   b) Mejorar la velocidad de carga del sitio web y asegurar que sea *responsive* y fácil de usar en dispositivos móviles.
   c) Incluir una densidad alta de palabras clave para asegurar que el sitio web aparezca en tantas búsquedas como sea posible.
   d) Obtener enlaces entrantes de otros sitios web de alta autoridad y relevancia temática.

6. Se puede afirmar que:
   a) No es posible identificar subgrupos dentro de una red utilizando análisis de redes.
   b) La cantidad de conexiones de un nodo siempre indica su influencia en la red.
   c) El tamaño de una red es siempre indicativo de su efectividad en la transmisión de información.
   d) Todas las relaciones en una red tienen la misma importancia para el análisis.

7. ¿Qué algoritmos permiten obtener información de una red?
   a) Índices de centralidad.
   b) Detección de comunidades.
   c) Hypertext Induced Topic Selection (HITS).
   d) PageRank.

8. El concepto de *rank sink* en el algoritmo de PageRank representa páginas web:
   a) Con distribuciones uniformes de la puntuación de PageRank.
   b) Que tienen una puntuación de PageRank más alta que otras debido a la manipulación de enlaces entrantes y salientes.
   c) Con una baja calidad de contenido y una cantidad insuficiente de enlaces salientes, lo que las hace menos relevantes en los resultados de búsqueda.
   d) Con un alto número de enlaces salientes que no reciben enlaces entrantes, lo que puede afectar negativamente su puntuación de PageRank.

9. En un sistema de control de tráfico urbano basado en reglas, ¿cuál de las siguientes reglas sería más efectiva para manejar situaciones de congestión en una intersección?

a) Si hay pocos vehículos en la intersección, reducir el tiempo de los semáforos en verde.

b) Si hay muchos vehículos en la intersección, aumentar el tiempo de los semáforos en verde.

c) Si hay un vehículo de emergencia en la intersección, detener todos los demás vehículos.

d) Si hay muchos peatones cruzando la intersección, reducir el tiempo de los semáforos en rojo.

10. Existen varias estrategias consideradas como no recomendables para mejorar el SEO de un sitio web. Dentro de estas estrategias negativas, se encuentran:

a) Construir enlaces naturales de sitios web con autoridad y relevancia temática.

b) Evitar el uso de las meta etiquetas y descripciones del sitio web para reflejar el contenido de la página.

c) Incluir una cantidad excesiva de palabras clave irrelevantes para intentar manipular los rankings de búsqueda.

d) Utilizar técnicas de encubrimiento para mostrar contenido diferente a lo indexado por los motores de búsqueda.

11. Con respecto a WordNet y su contribución a las ontologías y la representación del conocimiento en los sistemas de RI, puede afirmarse que:

a) En WordNet, una palabra está asociada a un conjunto de sinónimos (synsets), siendo estas palabras intercambiables en un contexto.

b) La integración de WordNet en sistemas de RI permite la expansión de consultas y la mejora de la precisión de los resultados al entender mejor el significado de los términos de búsqueda a través de su contexto semántico y las relaciones entre palabras.

c) WordNet diferencia claramente entre los significados de palabras según su uso en diferentes contextos, lo que permite aplicaciones avanzadas en desambiguación semántica más allá de los sistemas de recomendación y RI.

d) Una limitación de WordNet en la representación del conocimiento es su enfoque en el idioma inglés, lo que plantea desafíos en la aplicación global y la interoperabilidad con sistemas de información multilingües.

12. En el contexto del procesamiento de grandes conjuntos de datos, se puede asegurar que:

a) MapReduce es un enfoque para procesar datos de forma secuencial en un solo servidor para evitar problemas de concurrencia.

b) MapReduce divide una tarea en múltiples pasos de map y reduce que se ejecutan de forma secuencial en diferentes servidores para mejorar el rendimiento y la escalabilidad.

c) MapReduce no es adecuado para el procesamiento de datos no estructurados.

d) MapReduce solo puede manejar pequeñas cantidades de datos y no escala bien a grandes conjuntos de datos.

13. La integración de Hadoop y MapReduce en la RI trae como ventaja:

a) La eliminación de la necesidad de sistemas de bases de datos.

b) La posibilidad de la extracción de información relevante y la generación de resultados significativos de grandes conjuntos de datos.

c) La garantía de la privacidad absoluta de los datos procesados.

d) La reducción de los costos operativos a cero.

14. En el diseño de un videojuego de roles (RPG) en el que los personajes tienen atributos como salud, fuerza y velocidad, ¿cuál de las siguientes opciones representa mejor una implementación de la representación del conocimiento orientado a objetos?

a) Cada personaje se representa como una lista de cadenas de texto que describen sus características físicas y habilidades.

b) Cada personaje se representa como una función que calcula sus atributos en función de su nivel y experiencia.

c) Cada personaje se representa como una matriz de números que almacena sus valores de atributos.

d) Cada personaje se representa como un objeto con propiedades como salud, fuerza y velocidad, y métodos para modificar y consultar estos valores.

15. En el contexto de la RI en redes sociales se puede afirmar que:

a) El análisis de redes sociales se centra exclusivamente en la cantidad de seguidores que tiene un usuario en particular para determinar su influencia en la red.

b) El análisis de centralidad de intermediación se utiliza para identificar usuarios que son importantes en una red social debido a su posición como "puentes" entre diferentes grupos de usuarios.

c) El análisis de sentimientos se utiliza para determinar la popularidad de una publicación en redes sociales sin tener en cuenta la opinión de los usuarios.

d) El análisis de redes sociales no es útil para comprender la difusión de información en una red social específica, para ello se utiliza la medida de centralidad de vector propio.

16. Si se tiene el conjunto de páginas interconectadas $\{A \to B, C; B \to C; C \to A\}$, entonces la página con valor más alto de PageRank es:

a) A y B.

b) A.

c) C.

d) B.

17. El posicionamiento de un sitio web en los motores de búsqueda puede ser afectado por:

    a) Utilizar técnicas de encubrimiento para mostrar contenido diferente a lo indexado por los motores de búsqueda.

    b) Contenido sin valor en el sitio web.

    c) Mantener una estructura de URL clara y coherente.

    d) Obtener enlaces de sitios web irrelevantes y de baja calidad.

18. Referente al Web Crawler puede afirmarse que:

    a) Los hipervínculos encontrados en cada sitio web que no pertenecen al dominio donde fueron encontrados se desechan, puesto que no expande el conjunto de URLs sin visitar.

    b) Las páginas visitadas no se procesan nunca más.

    c) No necesita de un conjunto inicial de URLs para recorrer la Web.

    d) No tiene como objetivo indexar y recopilar información de diferentes sitios web.

19. En una red de transporte donde cada nodo es una parada de autobús y las aristas representan si existe un carro que pasa por ambos sitios, ¿qué puede mejorar el sistema de transporte?

    a) No considerar la centralidad de intermediación de las estaciones de transporte público al planificar rutas y horarios, ya que no tiene impacto en las conexiones entre las paradas.

    b) Utilizar el grafo inducido de los nodos con mayor valor en la centralidad de grado para aplicar la centralidad de intermediación con el propósito de reforzar las paradas con mayor tráfico.

    c) Utilizar la centralidad de intermediación para identificar las paradas de transferencia clave y establecer nuevas conexiones entre diferentes líneas de transporte público.

    d) No implementar sistemas de información en tiempo real para los usuarios, ya que pueden aumentar la carga de los trabajadores.

20. Sobre el algoritmo de PageRank, visto en clase, se puede afirmar que:

    a) Evalúa la importancia de un sitio web en función de la calidad y cantidad de enlaces entrantes que recibe de otros sitios web.

    b) Solo tiene en cuenta el contenido en un sitio web para determinar su relevancia en los resultados de búsqueda.

    c) Asigna una puntuación alta a los sitios web que tienen un gran número de enlaces entrantes sin tener en cuenta la calidad de esos enlaces.

    d) Asigna una puntuación baja a los sitios web que contienen muchos enlaces salientes, ya que indica una falta de relevancia.

Trabajo de Control
Sistemas de Recuperación de Información
Fecha: 15 de julio de 2024
Temario #51

Nombre: _____ Grupo: _____

Seleccione en cada caso la(s) respuesta(s) correcta(s). Considere también no marcar ninguna opción.

1. La diferencia entre la indexación por tokens y la indexación por conceptos puede definirse como:
   a) La indexación por tokens asigna pesos a los términos basados en su importancia relativa, mientras que la indexación por conceptos utiliza un sistema de etiquetado para asociar términos con características generales.
   b) La indexación por tokens divide los datos en términos individuales, mientras que la indexación por conceptos agrupa los datos en categorías definidas.
   c) La indexación por tokens asigna un valor numérico a cada término de los datos, mientras que la indexación por conceptos utiliza algoritmos de encriptación para proteger la privacidad de los datos.
   d) La indexación por tokens normaliza los datos reduciéndolos a su forma básica, mientras que la indexación por conceptos utiliza un método de ordenación para organizar los términos característicos de los datos.

2. En una biblioteca digital se necesita implementar un sistema que permita a los usuarios encontrar libros y artículos científicos de forma rápida y precisa. Los documentos están en diversos formatos, incluyendo PDF, EPUB y HTML. Se requiere seleccionar un algoritmo de indexación adecuado para el sistema, por lo que el programador designado para la implementación debe considerar:
   a) La velocidad de indexación y recuperación de datos.
   b) La capacidad para manejar documentos en diferentes formatos.
   c) La complejidad del algoritmo en términos de implementación y mantenimiento.
   d) La capacidad del algoritmo para procesar imágenes incrustadas o referenciadas en los ficheros.

3. En un grafo una comunidad es:
   a) Un conjunto de nodos que no comparten ninguna similitud estructural o funcional entre sí.
   b) Un conjunto de nodos aislados.
   c) Un conjunto de nodos altamente conectados que forman un subgrafo completamente independiente del resto de la red.
   d) Un conjunto de nodos que están más densamente interconectados entre sí que con los nodos fuera del conjunto.

4. El componente responsable de la gestión de recursos y planificación de tareas en Hadoop es:
   a) MapReduce.
   b) Hadoop Common.
   c) HDFS.
   d) YARN.

5. El algoritmo de PageRank converge si:
   a) La norma de la diferencia entre los vectores es menor a un umbral predefinido.
   b) Finaliza la ejecución cuando el número de iteraciones excede un máximo de iteraciones previamente definido.
   c) Se define un factor de normalización en la fórmula de la función.
   d) El algoritmo no se implementa de forma iterativa.

6. Un SRI es capaz de:
   a) Crear los índices asociados a los datos sin tener que analizar cada dato.
   b) Generar índices invertidos de manera óptima sin considerar el contexto.
   c) Reducir el tiempo de indexación de los datos si utiliza servidores distribuidos dentro de la red para que cada uno ejecute la indexación del mismo conjunto de datos.
   d) No necesitar de ningún almacenamiento externo para alojar los índices de los datos.

7. Con respecto a WordNet y su contribución a las ontologías y la representación del conocimiento en los sistemas de RI, puede afirmarse que:
   a) En WordNet, una palabra está asociada a un conjunto de sinónimos (synsets), siendo estas palabras intercambiables en un contexto.
   b) La integración de WordNet en sistemas de RI permite la expansión de consultas y la mejora de la precisión de los resultados al entender mejor el significado de los términos de búsqueda a través de su contexto semántico y las relaciones entre palabras.

c) WordNet diferencia claramente entre los significados de palabras según su uso en diferentes contextos, lo que permite aplicaciones avanzadas en desambiguación semántica más allá de los sistemas de recomendación y RI.

d) Una limitación de WordNet en la representación del conocimiento es su enfoque en el idioma inglés, lo que plantea desafíos en la aplicación global y la interoperabilidad con sistemas de información multilingües.

8. El concepto de *rank sink* en el algoritmo de PageRank representa páginas web:

a) Con distribuciones uniformes de la puntuación de PageRank.

b) Que tienen una puntuación de PageRank más alta que otras debido a la manipulación de enlaces entrantes y salientes.

c) Con una baja calidad de contenido y una cantidad insuficiente de enlaces salientes, lo que las hace menos relevantes en los resultados de búsqueda.

d) Con un alto número de enlaces salientes que no reciben enlaces entrantes, lo que puede afectar negativamente su puntuación de PageRank.

9. No se considera como técnica para detectar comunidades en una red:

a) Analizar la mutualidad de los enlaces.

b) Usar el agrupamiento jerárquico.

c) Utilizar el algoritmo de K-Means.

d) Encontrar cliques de vértices de grado par.

10. El uso de Hadoop y MapReduce en el contexto de la RI tiene como objetivo:

a) Automatizar el mantenimiento de sistemas de bases de datos.

b) Facilitar el análisis en tiempo real de datos de redes sociales.

c) Mejorar la eficiencia energética en centros de datos.

d) Procesar y analizar grandes conjuntos de datos para la RI.

11. La Web 1.0 se caracteriza por:

a) Las actualizaciones del contenido de los sitios web se efectúa de forma manual dentro del código HTML

b) Los sitios web se centran en brindar información en lugar de facilitar la colaboración o participación de los usuarios.

c) Uso de comunidades virtuales para popularizar los sitios web de noticias.

d) Los propietarios de los sitios web proporcionan contenido de forma periódica.

12. Se tiene un grafo donde cada nodo es un personaje de cierto libro de cuentos y la existencia de las aristas está definida si dos personajes aparecen en el mismo cuento. Se puede asegurar que:

a) La centralidad de intermediación identifica a los personajes que actúan como conectores entre personajes de cuentos distintos.

b) La centralidad de grado es útil para identificar los personajes con más conexiones dentro de la red, lo que puede indicar su importancia en el libro.

c) La centralidad de cercanía ofrece una relación entre la cantidad de vecinos de un nodo con respecto a la longitud máxima de un camino dentro del grafo partiendo del nodo en cuestión.

d) La centralidad de cercanía indica el grado de conexión de cada personaje con el resto de los personajes de los cuentos del libro.

13. Una buena práctica de SEO para mejorar el posicionamiento de un sitio web en los motores de búsqueda es:

a) Obtener enlaces de otros sitios web relevantes y de calidad que apunten al sitio.

b) Copiar contenido directamente de otros sitios web populares para aumentar la cantidad de páginas indexadas.

c) Utilizar técnicas de encubrimiento para mostrar contenido diferente a lo indexado por los motores de búsqueda.

d) Llenar el contenido con palabras clave irrelevantes para aumentar la densidad de palabras clave.

14. Se puede afirmar que:

a) No es posible identificar subgrupos dentro de una red utilizando análisis de redes.

b) La cantidad de conexiones de un nodo siempre indica su influencia en la red.

c) El tamaño de una red es siempre indicativo de su efectividad en la transmisión de información.

d) Todas las relaciones en una red tienen la misma importancia para el análisis.

15. Para contribuir positivamente al posicionamiento orgánico de un sitio web en los motores de búsqueda se puede:

a) Crear contenido relevante y de alta calidad que satisfaga las necesidades de información de los usuarios.

b) Mejorar la velocidad de carga del sitio web y asegurar que sea *responsive* y fácil de usar en dispositivos móviles.

c) Incluir una densidad alta de palabras clave para asegurar que el sitio web aparezca en tantas búsquedas como sea posible.

d) Obtener enlaces entrantes de otros sitios web de alta autoridad y relevancia temática.

16. En el contexto de la RI en redes, se puede afirmar que:
    a) La detección de comunidades no es relevante para la RI en redes, ya que se centra únicamente en la estructura de la red sin considerar el contenido.
    b) La detección de comunidades en una red siempre produce resultados objetivos y consistentes independientemente del algoritmo utilizado.
    c) La detección de comunidades ayuda a identificar grupos de nodos altamente conectados entre sí, lo que puede ser útil para comprender la estructura y el contenido de la red.
    d) La detección de comunidades solo se aplica a redes pequeñas y simples, no a redes grandes y complejas.

17. La transición de la Web 1.0 a la Web 2.0 se caracterizó principalmente por:
    a) El aumento en la velocidad de conexión a internet, que permitió una mejor calidad de las páginas web.
    b) La disminución de la importancia de los motores de búsqueda en la navegación web.
    c) El cambio de páginas web estáticas a dinámicas, permitiendo la interacción del usuario y la generación de contenido.
    d) La reducción en el uso de HTML y CSS en el desarrollo de sitios web.

18. En una empresa comercial, se tiene una red donde los nodos corresponden a los empleados y las aristas representan la frecuencia con la que colaboran en las ventas. Se busca mejorar la colaboración entre los empleados para aumentar las ventas totales, por lo que la directiva debe:
    a) Utilizar la centralidad de cercanía para identificar los empleados menos cercanos a otros en la red y asignarles tareas individuales para evitar posibles conflictos y desacuerdos en el proceso de colaboración.
    b) Utilizar la centralidad de cercanía para identificar los empleados más cercanos a otros en la red y promover la colaboración entre ellos, facilitando así la comunicación y el intercambio de conocimientos para mejorar las ventas.
    c) No considerar la centralidad de cercanía de los empleados al diseñar estrategias de colaboración, ya que todos los empleados tienen el mismo acceso a la información y los recursos.
    d) No tener en cuenta la centralidad de cercanía de los empleados al diseñar estrategias de colaboración, ya que esto podría introducir complicaciones adicionales en el proceso de trabajo.

19. Si una red cumple la propiedad de ser un grafo de mundo pequeño, entonces se conoce que:
    a) El número de componentes fuertemente conexas está relacionado con la cantidad de grafos K_n presentes.
    b) La longitud media del camino entre todo par de vértices es pequeña.
    c) El grafo es un anillo regular de grado 5.
    d) La red posee pocos vértices.

20. La Web 3 se conoce como:
    a) Internet de las cosas.
    b) Web de solo lectura.
    c) Web semántica.
    d) Web de escritura-lectura.

Trabajo de Control
Sistemas de Recuperación de Información
Fecha: 15 de julio de 2024
Temario #52

Nombre: _____ Grupo: _____

Seleccione en cada caso la(s) respuesta(s) correcta(s). Considere también no marcar ninguna opción.

1. Para la RI, el análisis de las redes puede:
   a) Ayudar a identificar grupos de interés.
   b) Indicar la importancia de una entidad en la transmisión de la información.
   c) Ayudar a comprender la conectividad y la accesibilidad entre las entidades.
   d) Revelar patrones de influencia dentro de una comunidad.

2. ¿Qué algoritmos permiten obtener información de una red?
   a) Índices de centralidad.
   b) Detección de comunidades.
   c) Hypertext Induced Topic Selection (HITS).
   d) PageRank.

3. La afirmación que mejor describe la política de amabilidad en los Web Crawlers es:
   a) Los crawlers se diseñan para acceder a sitios web sin restricciones y extraer datos de manera agresiva para su indexación.
   b) La política de amabilidad de los Web Crawlers dicta que los crawlers deben priorizar ciertos tipos de contenido sobre otros, ignorando completamente ciertas páginas web.
   c) La política de amabilidad establece pautas y reglas sobre cómo los crawlers deben interactuar con los sitios web para minimizar la carga del servidor y respetar las directivas de los administradores del sitio.
   d) Los Web Crawlers son libres de recopilar datos de cualquier sitio web sin restricciones, independientemente de la cantidad de tráfico que generen.

4. Los algoritmos para detectar comunidades en una red intentan:
   a) Buscar conjuntos donde cada nodo de un mismo conjunto tenga características similares al resto de los nodos del conjunto.
   b) Seleccionar aleatoriamente nodos de alto grado y sus vecinos.
   c) Buscar subgrafos tal que no incluyan nodos cuya ausencia desconecte al subgrafo.
   d) Encontrar grupos donde los nodos pertenecientes a los mismos grupos son cercanos bajo cierta métrica y lejanos con respecto a los nodos de otros grupos.

5. El uso de Hadoop y MapReduce en el contexto de la RI tiene como objetivo:
   a) Automatizar el mantenimiento de sistemas de bases de datos.
   b) Facilitar el análisis en tiempo real de datos de redes sociales.
   c) Mejorar la eficiencia energética en centros de datos.
   d) Procesar y analizar grandes conjuntos de datos para la RI.

6. Si se tiene el conjunto de páginas interconectadas $\{A \to B, C; B \to C; C \to A\}$, entonces la página con valor más alto de PageRank es:
   a) A y B.
   b) A.
   c) C.
   d) B.

7. La Web 3 se conoce como:
   a) Internet de las cosas.
   b) Web de solo lectura.
   c) Web semántica.
   d) Web de escritura-lectura.

8. En el contexto del procesamiento de grandes conjuntos de datos, se puede asegurar que:
   a) MapReduce es un enfoque para procesar datos de forma secuencial en un solo servidor para evitar problemas de concurrencia.
   b) MapReduce divide una tarea en múltiples pasos de map y reduce que se ejecutan de forma secuencial en diferentes servidores para mejorar el rendimiento y la escalabilidad.
   c) MapReduce no es adecuado para el procesamiento de datos no estructurados.
   d) MapReduce solo puede manejar pequeñas cantidades de datos y no escala bien a grandes conjuntos de datos.

9. Una buena práctica de SEO para mejorar el posicionamiento de un sitio web en los motores de búsqueda es:

a) Obtener enlaces de otros sitios web relevantes y de calidad que apunten al sitio.

b) Copiar contenido directamente de otros sitios web populares para aumentar la cantidad de páginas indexadas.

c) Utilizar técnicas de encubrimiento para mostrar contenido diferente a lo indexado por los motores de búsqueda.

d) Llenar el contenido con palabras clave irrelevantes para aumentar la densidad de palabras clave.

10. En un grafo una comunidad es:

a) Un conjunto de nodos que no comparten ninguna similitud estructural o funcional entre sí.

b) Un conjunto de nodos aislados.

c) Un conjunto de nodos altamente conectados que forman un subgrafo completamente independiente del resto de la red.

d) Un conjunto de nodos que están más densamente interconectados entre sí que con los nodos fuera del conjunto.

11. La premisa básica del algoritmo de PageRank para clasificar páginas web en los resultados de búsqueda es:

a) La antigüedad de la página web es el principal factor para determinar su clasificación.

b) El contenido y la relevancia de las palabras clave en la página web determinan su posición.

c) Los enlaces entrantes a una página web desde otras páginas contribuyen a su importancia y clasificación.

d) La cantidad de visitas que recibe una página web determina su clasificación.

12. En un sistema donde el conocimiento está orientado a objetos se puede asegurar que:

a) Se enfatiza la atención a la información de la cual se extrajo el conocimiento.

b) Los marcos y las bandas son las estructuras utilizadas para la representación del modelo.

c) No es posible definir especializaciones de los objetos de la vida real dentro del sistema.

d) Las funciones de agregación dificultan poder establecer relaciones entre los objetos.

13. Considerando las dimensiones y desafíos inherentes a Big Data puede afirmarse, tomando en cuenta las características clave y las implicaciones para su procesamiento y análisis, que:

a) Big Data se caracteriza principalmente por su pequeño volumen y uniformidad, permitiendo un procesamiento eficiente con mínimas adaptaciones de las herramientas de análisis de datos tradicionales.

b) Big Data no desafía la capacidad de las herramientas tradicionales de procesamiento de datos para capturar, almacenar, gestionar y analizar efectivamente la información, dada la evolución constante de las capacidades computacionales y algoritmos de optimización.

c) Aunque Big Data puede incluir datos estructurados, su naturaleza se expande al incorporar grandes cantidades de datos no estructurados y semiestructurados, lo que exige el uso de tecnologías especializadas en almacenamiento y procesamiento como Hadoop y sistemas de bases de datos NoSQL.

d) Además de su complejidad y diversidad, Big Data introduce desafíos significativos en términos de veracidad y variabilidad al requerir métodos avanzados de limpieza y validación de datos para asegurar la integridad del análisis.

14. Una empresa de comercio electrónico necesita mejorar su motor de búsqueda para proporcionar resultados más relevantes a sus usuarios. Actualmente, los resultados de la búsqueda no son precisos y los usuarios a menudo encuentran dificultades para encontrar productos específicos. La empresa está considerando implementar una indexación por conceptos para mejorar la relevancia de los resultados de búsqueda. ¿Qué beneficios podría provocar este cambio?

a) Permite adaptarse fácilmente a cambios en el vocabulario y la terminología utilizada en las descripciones de los productos.

b) Ayuda a identificar automáticamente relaciones entre productos, lo que puede mejorar las recomendaciones personalizadas a los usuarios.

c) Facilita la visualización de productos al agruparlos por categorías o características comunes.

d) Mejora la precisión en la búsqueda de productos relacionados, incluso cuando no coinciden exactamente con los términos de búsqueda del usuario.

15. El algoritmo de PageRank converge si:

a) La norma de la diferencia entre los vectores es menor a un umbral predefinido.

b) Finaliza la ejecución cuando el número de iteraciones excede un máximo de iteraciones previamente definido.

c) Se define un factor de normalización en la fórmula de la función.

d) El algoritmo no se implementa de forma iterativa.

16. En un SRI la indexación:

a) Consiste en asociar un identificador único a cada dato almacenado en el sistema.

b) Mejora la experiencia del usuario.

c) Permite la organización y la categorización de la información.

d) Facilita la RI relevante.

17. Para contribuir positivamente al posicionamiento orgánico de un sitio web en los motores de búsqueda se puede:
    a) Crear contenido relevante y de alta calidad que satisfaga las necesidades de información de los usuarios.
    b) Mejorar la velocidad de carga del sitio web y asegurar que sea *responsive* y fácil de usar en dispositivos móviles.
    c) Incluir una densidad alta de palabras clave para asegurar que el sitio web aparezca en tantas búsquedas como sea posible.
    d) Obtener enlaces entrantes de otros sitios web de alta autoridad y relevancia temática.
18. En el diseño de un videojuego de roles (RPG) en el que los personajes tienen atributos como salud, fuerza y velocidad, ¿cuál de las siguientes opciones representa mejor una implementación de la representación del conocimiento orientado a objetos?
    a) Cada personaje se representa como una lista de cadenas de texto que describen sus características físicas y habilidades.
    b) Cada personaje se representa como una función que calcula sus atributos en función de su nivel y experiencia.
    c) Cada personaje se representa como una matriz de números que almacena sus valores de atributos.
    d) Cada personaje se representa como un objeto con propiedades como salud, fuerza y velocidad, y métodos para modificar y consultar estos valores.
19. La web actual se enfrenta a problemas como:
    a) Presencia de grandes volúmenes de datos estructurados.
    b) Presencia de una alta calidad en los datos.
    c) Presencia de datos volátiles y distribuidos.
    d) Heterogeneidad en los datos.
20. Se puede afirmar que:
    a) El término "Big Data" se refiere exclusivamente al volumen de datos que una organización maneja, sin tener en cuenta la velocidad, la variedad y la veracidad de los datos.
    b) Los SRI pueden beneficiarse de MapReduce para mejorar la RI relevante.
    c) MapReduce es un modelo de procesamiento distribuido utilizado para trabajar con grandes volúmenes de datos.
    d) Uno de los desafíos en el procesamiento de Big Data es la capacidad de gestionar y analizar datos provenientes de diversas fuentes y en diferentes formatos de manera eficiente.

Trabajo de Control
Sistemas de Recuperación de Información
Fecha: 15 de julio de 2024
Temario #53

Nombre: _____ Grupo: _____

Seleccione en cada caso la(s) respuesta(s) correcta(s). Considere también no marcar ninguna opción.

1. En un grafo una comunidad es:
   a) Un conjunto de nodos que no comparten ninguna similitud estructural o funcional entre sí.
   b) Un conjunto de nodos aislados.
   c) Un conjunto de nodos altamente conectados que forman un subgrafo completamente independiente del resto de la red.
   d) Un conjunto de nodos que están más densamente interconectados entre sí que con los nodos fuera del conjunto.

2. En el contexto del procesamiento de grandes conjuntos de datos, se puede asegurar que:
   a) MapReduce es un enfoque para procesar datos de forma secuencial en un solo servidor para evitar problemas de concurrencia.
   b) MapReduce divide una tarea en múltiples pasos de map y reduce que se ejecutan de forma secuencial en diferentes servidores para mejorar el rendimiento y la escalabilidad.
   c) MapReduce no es adecuado para el procesamiento de datos no estructurados.
   d) MapReduce solo puede manejar pequeñas cantidades de datos y no escala bien a grandes conjuntos de datos.

3. Considerando las dimensiones y desafíos inherentes a Big Data puede afirmarse, tomando en cuenta las características clave y las implicaciones para su procesamiento y análisis, que:
   a) Big Data se caracteriza principalmente por su pequeño volumen y uniformidad, permitiendo un procesamiento eficiente con mínimas adaptaciones de las herramientas de análisis de datos tradicionales.
   b) Big Data no desafía la capacidad de las herramientas tradicionales de procesamiento de datos para capturar, almacenar, gestionar y analizar efectivamente la información, dada la evolución constante de las capacidades computacionales y algoritmos de optimización.
   c) Aunque Big Data puede incluir datos estructurados, su naturaleza se expande al incorporar grandes cantidades de datos no estructurados y semiestructurados, lo que exige el uso de tecnologías especializadas en almacenamiento y procesamiento como Hadoop y sistemas de bases de datos NoSQL.
   d) Además de su complejidad y diversidad, Big Data introduce desafíos significativos en términos de veracidad y variabilidad al requerir métodos avanzados de limpieza y validación de datos para asegurar la integridad del análisis.

4. En el modelo de representación del conocimiento basado en herencia se puede asegurar que:
   a) La cantidad de padres que puede tener un nodo no es mayor que 4.
   b) Las conclusiones pueden ser canceladas si el grafo es ambiguo.
   c) Solo se usa en entornos referentes a la biología.
   d) El razonamiento deducido está respaldado por al menos un camino dentro del grafo.

5. El algoritmo Hypertext Induced Topic Selection (HITS) intenta buscar nodos especiales. Estos son conocidos como:
   a) Autoridades.
   b) Hubs.
   c) Centrales.
   d) Sensibles.

6. Un sistema cuenta con la siguiente información:
   - Juan pertenece a la marina.
   - Juan es capellán (sacerdote encargado de una tarea específica fuera de la parroquia).
   - Los infantes de la marina suelen ser bebedores de cerveza.
   - Un capellán no suele ser bebedor de cerveza.
   - Un bebedor de cerveza suele tener sobrepeso.
   - Por lo general, un infante de la marina está en buena condición física.

   Si se representa la información en una red de herencia se puede concluir que:
   a) No se puede asegurar que Juan sea capellán.
   b) No se puede asegurar que Juan esté en buena condición física.
   c) Juan tiene sobrepeso.
   d) Existen dos razonamientos cancelables.

7. Se tiene una red de ingredientes donde cada uno representa un nodo y las aristas simbolizan que los ingredientes forman parte de una misma receta. Se busca mejorar la experiencia culinaria mediante la elaboración de combinaciones de ingredientes más interesantes y creativas, para lo cual se debe:
   a) No considerar la centralidad de grado de los ingredientes, ya que todas las combinaciones de ingredientes son igualmente válidas.
   b) Utilizar la centralidad de grado para identificar los ingredientes menos conectados en la red y tomarlos en cuenta para su inclusión en las combinaciones.
   c) No tener en cuenta la centralidad de grado de los ingredientes para no darle mayor importancia a los ingredientes más comunes.
   d) Utilizar la centralidad de grado para identificar los ingredientes más populares en la red y crear combinaciones que incluyan una variedad de ingredientes menos comunes.

8. Sobre el algoritmo de PageRank, visto en clase, se puede afirmar que:
   a) Evalúa la importancia de un sitio web en función de la calidad y cantidad de enlaces entrantes que recibe de otros sitios web.
   b) Solo tiene en cuenta el contenido en un sitio web para determinar su relevancia en los resultados de búsqueda.
   c) Asigna una puntuación alta a los sitios web que tienen un gran número de enlaces entrantes sin tener en cuenta la calidad de esos enlaces.
   d) Asigna una puntuación baja a los sitios web que contienen muchos enlaces salientes, ya que indica una falta de relevancia.

9. La afirmación que mejor describe la política de amabilidad en los Web Crawlers es:
   a) Los crawlers se diseñan para acceder a sitios web sin restricciones y extraer datos de manera agresiva para su indexación.
   b) La política de amabilidad de los Web Crawlers dicta que los crawlers deben priorizar ciertos tipos de contenido sobre otros, ignorando completamente ciertas páginas web.
   c) La política de amabilidad establece pautas y reglas sobre cómo los crawlers deben interactuar con los sitios web para minimizar la carga del servidor y respetar las directivas de los administradores del sitio.
   d) Los Web Crawlers son libres de recopilar datos de cualquier sitio web sin restricciones, independientemente de la cantidad de tráfico que generen.

10. El conjunto frontera de URLs en un Web Crawler:
   a) Delimita los sitios web que visitará en futuras iteraciones del proceso.
   b) Es similar a un conjunto de URLs que esperan ser visitadas.
   c) Almacena hipervínculos que pertenecen al mismo dominio del conjunto semilla de URLs con que inició el crawler.
   d) Indica las secciones que no pueden ser visitadas de cada sitio web.

11. La computación evolutiva:
   a) No es aplicable en la RI debido a la complejidad de los algoritmos evolutivos.
   b) Solo puede manejar conjuntos de datos pequeños y no es escalable a grandes volúmenes de datos.
   c) Utiliza algoritmos para buscar soluciones óptimas en grandes espacios de búsqueda, lo que la hace adecuada para problemas de optimización en la RI.
   d) Es útil solo para problemas de clasificación de documentos y no para otras tareas de RI en general.

12. En el contexto de la RI en redes sociales se puede afirmar que:
   a) El análisis de redes sociales se centra exclusivamente en la cantidad de seguidores que tiene un usuario en particular para determinar su influencia en la red.
   b) El análisis de centralidad de intermediación se utiliza para identificar usuarios que son importantes en una red social debido a su posición como "puentes" entre diferentes grupos de usuarios.
   c) El análisis de sentimientos se utiliza para determinar la popularidad de una publicación en redes sociales sin tener en cuenta la opinión de los usuarios.
   d) El análisis de redes sociales no es útil para comprender la difusión de información en una red social específica, para ello se utiliza la medida de centralidad de vector propio.

13. En la RI en el contexto de Big Data, se puede asegurar que:
   a) El almacenamiento distribuido centraliza todos los datos en un único servidor para facilitar su gestión y mantenimiento.
   b) El almacenamiento distribuido ofrece ventajas significativas en términos de escalabilidad y rendimiento en comparación con el almacenamiento centralizado.
   c) El almacenamiento distribuido reparte los datos en múltiples servidores para mejorar la disponibilidad y la redundancia del sistema.
   d) El almacenamiento distribuido es una técnica obsoleta.

14. En la RI en el contexto de Big Data, se puede asegurar que:

a) La indexación distribuida divide los datos en múltiples fragmentos que se almacenan en varios nodos para permitir búsquedas paralelas y mejorar la escalabilidad.

b) La indexación distribuida no ofrece ventajas en términos de rendimiento y escalabilidad en comparación con la indexación centralizada.

c) La indexación distribuida es una técnica obsoleta.

d) La indexación distribuida almacena todos los datos en un solo servidor para facilitar su acceso y búsqueda.

15. Al realizar la optimización de contenido para SEO debe considerarse:

a) Seleccionar las palabras clave al azar.

b) Utilizar etiquetas de título y meta descripciones únicas y relevantes para cada página.

c) Incluir palabras clave de manera excesiva en el contenido para mejorar el posicionamiento.

d) Crear contenido valioso y original que satisfaga las necesidades de los usuarios.

16. En una biblioteca digital se necesita implementar un sistema que permita a los usuarios encontrar libros y artículos científicos de forma rápida y precisa. Los documentos están en diversos formatos, incluyendo PDF, EPUB y HTML. Se requiere seleccionar un algoritmo de indexación adecuado para el sistema, por lo que el programador designado para la implementación debe considerar:

a) La velocidad de indexación y recuperación de datos.

b) La capacidad para manejar documentos en diferentes formatos.

c) La complejidad del algoritmo en términos de implementación y mantenimiento.

d) La capacidad del algoritmo para procesar imágenes incrustadas o referenciadas en los ficheros.

17. No se considera como técnica para detectar comunidades en una red:

a) Analizar la mutualidad de los enlaces.

b) Usar el agrupamiento jerárquico.

c) Utilizar el algoritmo de K-Means.

d) Encontrar cliques de vértices de grado par.

18. Un investigador necesita recopilar datos de múltiples sitios web para un estudio académico, pero se enfrenta a varios desafíos al realizar el proceso de Web Scraping de manera ética y legal. ¿Cuál de las siguientes opciones describe mejor uno de los desafíos asociados al proceso de Web Scraping?

a) La necesidad de comprender la estructura del sitio web y su código HTML para extraer los datos correctamente.

b) La disponibilidad limitada de datos en línea que se pueden extraer utilizando técnicas de Web Scraping.

c) La dificultad para encontrar herramientas de Web Scraping gratuitas y fiables.

d) La necesidad de estar montado sobre un scrawler que cumpla con todas las políticas.

19. Si se tiene el conjunto de páginas interconectadas $\{A \to B, C; B \to C; C \to A\}$, entonces la página con valor más alto de PageRank es:

a) A y B.

b) A.

c) C.

d) B.

20. Para transformar el contenido no estructurado de las páginas web en datos estructurados, el Web Scraping:

a) Extrae información basada en patrones de HTML/CSS.

b) Analiza los protocolos de red.

c) Interpreta el código JavaScript en tiempo real.

d) Convierte de forma automática imágenes a texto.

Trabajo de Control
Sistemas de Recuperación de Información
Fecha: 15 de julio de 2024
Temario #54

Nombre: _____ Grupo: _____

Seleccione en cada caso la(s) respuesta(s) correcta(s). Considere también no marcar ninguna opción.

1. Una plataforma de comercio electrónico desea mejorar la experiencia del usuario al permitir una navegación más personalizada y contextualizada. Actualmente los usuarios tienen dificultades para encontrar productos relevantes debido a la gran cantidad de opciones disponibles. La empresa está interesada en implementar características de la Web 2.5 y la Web Semántica para abordar este problema. ¿Qué características podrían ayudar para ofrecer una navegación más personalizada y contextualizada?
   a) La implementación de ontologías y metadatos para enriquecer la descripción de productos y mejorar la precisión de las recomendaciones.
   b) La optimización de la velocidad de carga del sitio web para mejorar la experiencia del usuario y reducir el abandono del carrito de compra.
   c) La integración de redes sociales para permitir la recomendación de productos basada en las preferencias de amigos y contactos.
   d) La incorporación de los productos en tendencia en el mercado.

2. Se puede afirmar que:
   a) El término "Big Data" se refiere exclusivamente al volumen de datos que una organización maneja, sin tener en cuenta la velocidad, la variedad y la veracidad de los datos.
   b) Los SRI pueden beneficiarse de MapReduce para mejorar la RI relevante.
   c) MapReduce es un modelo de procesamiento distribuido utilizado para trabajar con grandes volúmenes de datos.
   d) Uno de los desafíos en el procesamiento de Big Data es la capacidad de gestionar y analizar datos provenientes de diversas fuentes y en diferentes formatos de manera eficiente.

3. Sobre el algoritmo de PageRank, visto en clase, se puede afirmar que:
   a) Evalúa la importancia de un sitio web en función de la calidad y cantidad de enlaces entrantes que recibe de otros sitios web.
   b) Solo tiene en cuenta el contenido en un sitio web para determinar su relevancia en los resultados de búsqueda.
   c) Asigna una puntuación alta a los sitios web que tienen un gran número de enlaces entrantes sin tener en cuenta la calidad de esos enlaces.
   d) Asigna una puntuación baja a los sitios web que contienen muchos enlaces salientes, ya que indica una falta de relevancia.

4. Dentro del análisis de redes, la centralidad de grado mide la importancia de un nodo basándose en:
   a) El grado del nodo.
   b) La cantidad de veces que aparece el nodo en el camino mínimo entre cualquier par de nodos.
   c) La cantidad de vecinos del nodo.
   d) El número de aristas que posee el nodo.

5. Implementar índices invertidos en un SRI asegura:
   a) Acelerar el proceso de búsqueda al permitir búsquedas directas por contenido en lugar de por título.
   b) Reducir la cantidad de espacio de almacenamiento necesario al comprimir los datos de los documentos.
   c) Facilitar la búsqueda sobre los datos que contienen términos específicos al mantener una lista de datos para cada término único.
   d) Incrementar la seguridad de los datos almacenados al dificultar el acceso directo a la información sin el índice correcto.

6. En un grafo una comunidad es:
   a) Un conjunto de nodos que no comparten ninguna similitud estructural o funcional entre sí.
   b) Un conjunto de nodos aislados.
   c) Un conjunto de nodos altamente conectados que forman un subgrafo completamente independiente del resto de la red.
   d) Un conjunto de nodos que están más densamente interconectados entre sí que con los nodos fuera del conjunto.

7. Se tiene un grafo donde cada nodo es un personaje de cierto libro de cuentos y la existencia de las aristas está definida si dos personajes aparecen en el mismo cuento. Se puede asegurar que:
   a) La centralidad de intermediación identifica a los personajes que actúan como conectores entre personajes de cuentos distintos.

b) La centralidad de grado es útil para identificar los personajes con más conexiones dentro de la red, lo que puede indicar su importancia en el libro.

c) La centralidad de cercanía ofrece una relación entre la cantidad de vecinos de un nodo con respecto a la longitud máxima de un camino dentro del grafo partiendo del nodo en cuestión.

d) La centralidad de cercanía indica el grado de conexión de cada personaje con el resto de los personajes de los cuentos del libro.

8. Para la RI, el análisis de las redes puede:
   a) Ayudar a identificar grupos de interés.
   b) Indicar la importancia de una entidad en la transmisión de la información.
   c) Ayudar a comprender la conectividad y la accesibilidad entre las entidades.
   d) Revelar patrones de influencia dentro de una comunidad.

9. Existen varias estrategias consideradas como no recomendables para mejorar el SEO de un sitio web. Dentro de estas estrategias negativas, se encuentran:
   a) Construir enlaces naturales de sitios web con autoridad y relevancia temática.
   b) Evitar el uso de las meta etiquetas y descripciones del sitio web para reflejar el contenido de la página.
   c) Incluir una cantidad excesiva de palabras clave irrelevantes para intentar manipular los rankings de búsqueda.
   d) Utilizar técnicas de encubrimiento para mostrar contenido diferente a lo indexado por los motores de búsqueda.

10. El algoritmo de PageRank puede describirse como un procedimiento utilizado para:
    a) Determinar la velocidad de carga de una página web en un navegador.
    b) Calcular la relevancia de una página web en función de la cantidad y calidad de los enlaces que apuntan hacia ella.
    c) Prevenir el spam y el contenido no deseado en las páginas web.
    d) Clasificar las páginas web en función de su edad y autoridad.

11. Para transformar el contenido no estructurado de las páginas web en datos estructurados, el Web Scraping:
    a) Extrae información basada en patrones de HTML/CSS.
    b) Analiza los protocolos de red.
    c) Interpreta el código JavaScript en tiempo real.
    d) Convierte de forma automática imágenes a texto.

12. El algoritmo Hypertext Induced Topic Selection (HITS) intenta buscar nodos especiales. Estos son conocidos como:
    a) Autoridades.
    b) Hubs.
    c) Centrales.
    d) Sensibles.

13. Sobre WordNet y su aplicación en el procesamiento del lenguaje natural, se puede afirmar que:
    a) El diseño de WordNet facilita su integración en aplicaciones multilingües de NLP, aunque su desarrollo original se centró en el inglés.
    b) Aunque WordNet es una herramienta valiosa en el NLP, su estructura no incluye información sobre la frecuencia de uso de las palabras en el lenguaje natural.
    c) Los synsets facilitan la identificación de relaciones semánticas entre palabras, como la hiperonimia y la meronimia, enriqueciendo tareas de NLP.
    d) WordNet proporciona una base para la desambiguación semántica al agrupar palabras con significados similares en synsets.

14. Si una red cumple la propiedad de ser un grafo de mundo pequeño, entonces se conoce que:
    a) El número de componentes fuertemente conexas está relacionado con la cantidad de grafos K_n presentes.
    b) La longitud media del camino entre todo par de vértices es pequeña.
    c) El grafo es un anillo regular de grado 5.
    d) La red posee pocos vértices.

15. La afirmación que mejor refleja el principio subyacente de PageRank, considerando su importancia en la clasificación de los sitios web, es:
    a) PageRank valora más la cantidad de enlaces entrantes a una página web, independientemente de la calidad o relevancia de estos enlaces.
    b) La efectividad de PageRank se basa exclusivamente en el análisis de las palabras clave contenidas en los enlaces entrantes, sin considerar la estructura de enlace de la Web.
    c) El algoritmo de PageRank considera tanto la cantidad como la calidad de los enlaces entrantes, asignando mayor valor a los enlaces provenientes de sitios web considerados como "importantes".
    d) PageRank opera bajo el supuesto de que los enlaces entrantes y salientes tienen el mismo impacto en la valoración de la relevancia de una página web.

16. En un sistema donde el conocimiento está definido a partir de reglas se puede asegurar que:
    a) El nivel de especificidad de las reglas está limitado.

b) La representación de conocimiento está basada en la lógica proposicional.

c) El razonamiento se activa en cadena hacia delante.

d) El orden en que se definen las reglas no altera el razonamiento del sistema.

17. Analizar una red permite:

a) Detectar posibles tendencias antes de que se conviertan en tendencia.

b) Obtener predicciones exactas de eventos futuros en mercados financieros.

c) Evaluar la calidad del contenido de un sitio web.

d) Encontrar nodos "sensibles" o críticos para la red.

18. La computación evolutiva:

a) No es aplicable en la RI debido a la complejidad de los algoritmos evolutivos.

b) Solo puede manejar conjuntos de datos pequeños y no es escalable a grandes volúmenes de datos.

c) Utiliza algoritmos para buscar soluciones óptimas en grandes espacios de búsqueda, lo que la hace adecuada para problemas de optimización en la RI.

d) Es útil solo para problemas de clasificación de documentos y no para otras tareas de RI en general.

19. En la RI en el contexto de Big Data, se puede asegurar que:

a) El almacenamiento distribuido centraliza todos los datos en un único servidor para facilitar su gestión y mantenimiento.

b) El almacenamiento distribuido ofrece ventajas significativas en términos de escalabilidad y rendimiento en comparación con el almacenamiento centralizado.

c) El almacenamiento distribuido reparte los datos en múltiples servidores para mejorar la disponibilidad y la redundancia del sistema.

d) El almacenamiento distribuido es una técnica obsoleta.

20. Se tiene una red de ingredientes donde cada uno representa un nodo y las aristas simbolizan que los ingredientes forman parte de una misma receta. Se busca mejorar la experiencia culinaria mediante la elaboración de combinaciones de ingredientes más interesantes y creativas, para lo cual se debe:

a) No considerar la centralidad de grado de los ingredientes, ya que todas las combinaciones de ingredientes son igualmente válidas.

b) Utilizar la centralidad de grado para identificar los ingredientes menos conectados en la red y tomarlos en cuenta para su inclusión en las combinaciones.

c) No tener en cuenta la centralidad de grado de los ingredientes para no darle mayor importancia a los ingredientes más comunes.

d) Utilizar la centralidad de grado para identificar los ingredientes más populares en la red y crear combinaciones que incluyan una variedad de ingredientes menos comunes.

Trabajo de Control
Sistemas de Recuperación de Información
Fecha: 15 de julio de 2024
Temario #55

Nombre: _____ Grupo: _____

Seleccione en cada caso la(s) respuesta(s) correcta(s). Considere también no marcar ninguna opción.

1. Para transformar el contenido no estructurado de las páginas web en datos estructurados, el Web Scraping:
   a) Extrae información basada en patrones de HTML/CSS.
   b) Analiza los protocolos de red.
   c) Interpreta el código JavaScript en tiempo real.
   d) Convierte de forma automática imágenes a texto.

2. Al realizar la optimización de contenido para SEO debe considerarse:
   a) Seleccionar las palabras clave al azar.
   b) Utilizar etiquetas de título y meta descripciones únicas y relevantes para cada página.
   c) Incluir palabras clave de manera excesiva en el contenido para mejorar el posicionamiento.
   d) Crear contenido valioso y original que satisfaga las necesidades de los usuarios.

3. La afirmación que mejor refleja el principio subyacente de PageRank, considerando su importancia en la clasificación de los sitios web, es:
   a) PageRank valora más la cantidad de enlaces entrantes a una página web, independientemente de la calidad o relevancia de estos enlaces.
   b) La efectividad de PageRank se basa exclusivamente en el análisis de las palabras clave contenidas en los enlaces entrantes, sin considerar la estructura de enlace de la Web.
   c) El algoritmo de PageRank considera tanto la cantidad como la calidad de los enlaces entrantes, asignando mayor valor a los enlaces provenientes de sitios web considerados como "importantes".
   d) PageRank opera bajo el supuesto de que los enlaces entrantes y salientes tienen el mismo impacto en la valoración de la relevancia de una página web.

4. Un SRI es capaz de:
   a) Crear los índices asociados a los datos sin tener que analizar cada dato.
   b) Generar índices invertidos de manera óptima sin considerar el contexto.
   c) Reducir el tiempo de indexación de los datos si utiliza servidores distribuidos dentro de la red para que cada uno ejecute la indexación del mismo conjunto de datos.
   d) No necesitar de ningún almacenamiento externo para alojar los índices de los datos.

5. Referente al Web Crawler puede afirmarse que:
   a) Los hipervínculos encontrados en cada sitio web que no pertenecen al dominio donde fueron encontrados se desechan, puesto que no expande el conjunto de URLs sin visitar.
   b) Las páginas visitadas no se procesan nunca más.
   c) No necesita de un conjunto inicial de URLs para recorrer la Web.
   d) No tiene como objetivo indexar y recopilar información de diferentes sitios web.

6. Si se tiene el conjunto de páginas interconectadas $\{A \to B, C; B \to C; C \to A\}$, entonces la página con valor más alto de PageRank es:
   a) A y B.
   b) A.
   c) C.
   d) B.

7. Implementar índices invertidos en un SRI asegura:
   a) Acelerar el proceso de búsqueda al permitir búsquedas directas por contenido en lugar de por título.
   b) Reducir la cantidad de espacio de almacenamiento necesario al comprimir los datos de los documentos.
   c) Facilitar la búsqueda sobre los datos que contienen términos específicos al mantener una lista de datos para cada término único.
   d) Incrementar la seguridad de los datos almacenados al dificultar el acceso directo a la información sin el índice correcto.

8. Para contribuir positivamente al posicionamiento orgánico de un sitio web en los motores de búsqueda se puede:
   a) Crear contenido relevante y de alta calidad que satisfaga las necesidades de información de los usuarios.
   b) Mejorar la velocidad de carga del sitio web y asegurar que sea *responsive* y fácil de usar en dispositivos móviles.

c) Incluir una densidad alta de palabras clave para asegurar que el sitio web aparezca en tantas búsquedas como sea posible.

d) Obtener enlaces entrantes de otros sitios web de alta autoridad y relevancia temática.

9. La afirmación que mejor describe la política de amabilidad en los Web Crawlers es:

a) Los crawlers se diseñan para acceder a sitios web sin restricciones y extraer datos de manera agresiva para su indexación.

b) La política de amabilidad de los Web Crawlers dicta que los crawlers deben priorizar ciertos tipos de contenido sobre otros, ignorando completamente ciertas páginas web.

c) La política de amabilidad establece pautas y reglas sobre cómo los crawlers deben interactuar con los sitios web para minimizar la carga del servidor y respetar las directivas de los administradores del sitio.

d) Los Web Crawlers son libres de recopilar datos de cualquier sitio web sin restricciones, independientemente de la cantidad de tráfico que generen.

10. Se puede afirmar que:

a) El término "Big Data" se refiere exclusivamente al volumen de datos que una organización maneja, sin tener en cuenta la velocidad, la variedad y la veracidad de los datos.

b) Los SRI pueden beneficiarse de MapReduce para mejorar la RI relevante.

c) MapReduce es un modelo de procesamiento distribuido utilizado para trabajar con grandes volúmenes de datos.

d) Uno de los desafíos en el procesamiento de Big Data es la capacidad de gestionar y analizar datos provenientes de diversas fuentes y en diferentes formatos de manera eficiente.

11. Un investigador necesita recopilar datos de múltiples sitios web para un estudio académico, pero se enfrenta a varios desafíos al realizar el proceso de Web Scraping de manera ética y legal. ¿Cuál de las siguientes opciones describe mejor uno de los desafíos asociados al proceso de Web Scraping?

a) La necesidad de comprender la estructura del sitio web y su código HTML para extraer los datos correctamente.

b) La disponibilidad limitada de datos en línea que se pueden extraer utilizando técnicas de Web Scraping.

c) La dificultad para encontrar herramientas de Web Scraping gratuitas y fiables.

d) La necesidad de estar montado sobre un scrawler que cumpla con todas las políticas.

12. Con respecto a WordNet y su contribución a las ontologías y la representación del conocimiento en los sistemas de RI, puede afirmarse que:

a) En WordNet, una palabra está asociada a un conjunto de sinónimos (synsets), siendo estas palabras intercambiables en un contexto.

b) La integración de WordNet en sistemas de RI permite la expansión de consultas y la mejora de la precisión de los resultados al entender mejor el significado de los términos de búsqueda a través de su contexto semántico y las relaciones entre palabras.

c) WordNet diferencia claramente entre los significados de palabras según su uso en diferentes contextos, lo que permite aplicaciones avanzadas en desambiguación semántica más allá de los sistemas de recomendación y RI.

d) Una limitación de WordNet en la representación del conocimiento es su enfoque en el idioma inglés, lo que plantea desafíos en la aplicación global y la interoperabilidad con sistemas de información multilingües.

13. En el contexto de la representación del conocimiento basada en herencia, ¿qué caracteriza a la herencia cancelable?

a) La herencia es el resultado del razonamiento no transitivo.

b) Las conclusiones no están determinadas y dependen del nodo de interés.

c) Las propiedades heredadas siempre se mantienen y no pueden anularse.

d) No existe ambigüedad en las conclusiones obtenidas.

14. El algoritmo de PageRank converge si:

a) La norma de la diferencia entre los vectores es menor a un umbral predefinido.

b) Finaliza la ejecución cuando el número de iteraciones excede un máximo de iteraciones previamente definido.

c) Se define un factor de normalización en la fórmula de la función.

d) El algoritmo no se implementa de forma iterativa.

15. Analizar una red permite:

a) Detectar posibles tendencias antes de que se conviertan en tendencia.

b) Obtener predicciones exactas de eventos futuros en mercados financieros.

c) Evaluar la calidad del contenido de un sitio web.

d) Encontrar nodos "sensibles" o críticos para la red.

16. Una empresa de comercio electrónico necesita mejorar su motor de búsqueda para proporcionar resultados más relevantes a sus usuarios. Actualmente, los resultados de la búsqueda no son precisos y los usuarios a menudo encuentran dificultades para encontrar productos específicos. La empresa está considerando implementar una indexación por conceptos para mejorar la relevancia de los resultados de búsqueda. ¿Qué beneficios podría provocar este cambio?

    a) Permite adaptarse fácilmente a cambios en el vocabulario y la terminología utilizada en las descripciones de los productos.

    b) Ayuda a identificar automáticamente relaciones entre productos, lo que puede mejorar las recomendaciones personalizadas a los usuarios.

    c) Facilita la visualización de productos al agruparlos por categorías o características comunes.

    d) Mejora la precisión en la búsqueda de productos relacionados, incluso cuando no coinciden exactamente con los términos de búsqueda del usuario.

17. En una red de transporte donde cada nodo es una parada de autobús y las aristas representan si existe un carro que pasa por ambos sitios, ¿qué puede mejorar el sistema de transporte?

    a) No considerar la centralidad de intermediación de las estaciones de transporte público al planificar rutas y horarios, ya que no tiene impacto en las conexiones entre las paradas.

    b) Utilizar el grafo inducido de los nodos con mayor valor en la centralidad de grado para aplicar la centralidad de intermediación con el propósito de reforzar las paradas con mayor tráfico.

    c) Utilizar la centralidad de intermediación para identificar las paradas de transferencia clave y establecer nuevas conexiones entre diferentes líneas de transporte público.

    d) No implementar sistemas de información en tiempo real para los usuarios, ya que pueden aumentar la carga de los trabajadores.

18. Considerando las dimensiones y desafíos inherentes a Big Data puede afirmarse, tomando en cuenta las características clave y las implicaciones para su procesamiento y análisis, que:

    a) Big Data se caracteriza principalmente por su pequeño volumen y uniformidad, permitiendo un procesamiento eficiente con mínimas adaptaciones de las herramientas de análisis de datos tradicionales.

    b) Big Data no desafía la capacidad de las herramientas tradicionales de procesamiento de datos para capturar, almacenar, gestionar y analizar efectivamente la información, dada la evolución constante de las capacidades computacionales y algoritmos de optimización.

    c) Aunque Big Data puede incluir datos estructurados, su naturaleza se expande al incorporar grandes cantidades de datos no estructurados y semiestructurados, lo que exige el uso de tecnologías especializadas en almacenamiento y procesamiento como Hadoop y sistemas de bases de datos NoSQL.

    d) Además de su complejidad y diversidad, Big Data introduce desafíos significativos en términos de veracidad y variabilidad al requerir métodos avanzados de limpieza y validación de datos para asegurar la integridad del análisis.

19. Considerando las prácticas éticas y legales en el Web Scraping, se puede asegurar que:

    a) Web Scraping requiere considerar las políticas de `robots.txt` del sitio web objetivo.

    b) Es importante revisar y respetar los términos de servicio del sitio web, así como las leyes aplicables de protección de datos y derechos de autor, antes de realizar Web Scraping.

    c) Web Scraping sobre datos personales sin consentimiento es generalmente aceptado si los datos se utilizan con fines de investigación.

    d) La extracción de datos mediante Web Scraping siempre es legal, independientemente de las leyes locales sobre derechos de autor y privacidad de datos definidos en los sitios web.

20. El uso de Hadoop y MapReduce en el contexto de la RI tiene como objetivo:

    a) Automatizar el mantenimiento de sistemas de bases de datos.

    b) Facilitar el análisis en tiempo real de datos de redes sociales.

    c) Mejorar la eficiencia energética en centros de datos.

    d) Procesar y analizar grandes conjuntos de datos para la RI.

Trabajo de Control
Sistemas de Recuperación de Información
Fecha: 15 de julio de 2024
Temario #56

Nombre: _____ Grupo: _____

Seleccione en cada caso la(s) respuesta(s) correcta(s). Considere también no marcar ninguna opción.

1. El uso de Hadoop y MapReduce en el contexto de la RI tiene como objetivo:
   a) Automatizar el mantenimiento de sistemas de bases de datos.
   b) Facilitar el análisis en tiempo real de datos de redes sociales.
   c) Mejorar la eficiencia energética en centros de datos.
   d) Procesar y analizar grandes conjuntos de datos para la RI.

2. Al "relajar" el concepto de clique en la detección de comunidades se intenta solucionar:
   a) La imposibilidad de conectar nodos distantes.
   b) La necesidad de datos externos para analizar la red.
   c) La uniformidad de los nodos en términos de grado.
   d) El solapamiento y la complejidad computacional.

3. En el contexto de la representación del conocimiento basada en herencia, ¿qué caracteriza a la herencia cancelable?
   a) La herencia es el resultado del razonamiento no transitivo.
   b) Las conclusiones no están determinadas y dependen del nodo de interés.
   c) Las propiedades heredadas siempre se mantienen y no pueden anularse.
   d) No existe ambigüedad en las conclusiones obtenidas.

4. En el contexto de la RI en redes sociales se puede afirmar que:
   a) El análisis de redes sociales se centra exclusivamente en la cantidad de seguidores que tiene un usuario en particular para determinar su influencia en la red.
   b) El análisis de centralidad de intermediación se utiliza para identificar usuarios que son importantes en una red social debido a su posición como "puentes" entre diferentes grupos de usuarios.
   c) El análisis de sentimientos se utiliza para determinar la popularidad de una publicación en redes sociales sin tener en cuenta la opinión de los usuarios.
   d) El análisis de redes sociales no es útil para comprender la difusión de información en una red social específica, para ello se utiliza la medida de centralidad de vector propio.

5. En una red de transporte donde cada nodo es una parada de autobús y las aristas representan si existe un carro que pasa por ambos sitios, ¿qué puede mejorar el sistema de transporte?
   a) No considerar la centralidad de intermediación de las estaciones de transporte público al planificar rutas y horarios, ya que no tiene impacto en las conexiones entre las paradas.
   b) Utilizar el grafo inducido de los nodos con mayor valor en la centralidad de grado para aplicar la centralidad de intermediación con el propósito de reforzar las paradas con mayor tráfico.
   c) Utilizar la centralidad de intermediación para identificar las paradas de transferencia clave y establecer nuevas conexiones entre diferentes líneas de transporte público.
   d) No implementar sistemas de información en tiempo real para los usuarios, ya que pueden aumentar la carga de los trabajadores.

6. La afirmación que mejor describe la política de amabilidad en los Web Crawlers es:
   a) Los crawlers se diseñan para acceder a sitios web sin restricciones y extraer datos de manera agresiva para su indexación.
   b) La política de amabilidad de los Web Crawlers dicta que los crawlers deben priorizar ciertos tipos de contenido sobre otros, ignorando completamente ciertas páginas web.
   c) La política de amabilidad establece pautas y reglas sobre cómo los crawlers deben interactuar con los sitios web para minimizar la carga del servidor y respetar las directivas de los administradores del sitio.
   d) Los Web Crawlers son libres de recopilar datos de cualquier sitio web sin restricciones, independientemente de la cantidad de tráfico que generen.

7. La centralidad de intermediación de un nodo indica:
   a) La cantidad de nodos vecinos directos.
   b) La resistencia del nodo a fallos.
   c) El número total de conexiones entrantes y salientes.
   d) La frecuencia con la que un nodo actúa como puente en el camino más corto entre otros dos nodos.

8. El conjunto frontera de URLs en un Web Crawler:
   a) Delimita los sitios web que visitará en futuras iteraciones del proceso.
   b) Es similar a un conjunto de URLs que esperan ser visitadas.
   c) Almacena hipervínculos que pertenecen al mismo dominio del conjunto semilla de URLs con que inició el crawler.
   d) Indica las secciones que no pueden ser visitadas de cada sitio web.

9. La política de ordenación de URLs en los Web Crawlers tiene como aspecto fundamental:
   a) Limitar el acceso de los crawlers a ciertas secciones de un sitio web, evitando el rastreo de URLs consideradas menos importantes o sensibles.
   b) Definir la estructura de la URL de destino, asegurando que estén ordenadas alfabéticamente para facilitar la navegación y la indexación.
   c) Establecer la prioridad de rastreo de las URLs, determinando el orden en que los crawlers visitan y procesan cada página web.
   d) Determinar la forma en que los crawlers asignan un valor de relevancia a cada URL para clasificarlas en el índice de búsqueda.

10. Considerando las dimensiones y desafíos inherentes a Big Data puede afirmarse, tomando en cuenta las características clave y las implicaciones para su procesamiento y análisis, que:
    a) Big Data se caracteriza principalmente por su pequeño volumen y uniformidad, permitiendo un procesamiento eficiente con mínimas adaptaciones de las herramientas de análisis de datos tradicionales.
    b) Big Data no desafía la capacidad de las herramientas tradicionales de procesamiento de datos para capturar, almacenar, gestionar y analizar efectivamente la información, dada la evolución constante de las capacidades computacionales y algoritmos de optimización.
    c) Aunque Big Data puede incluir datos estructurados, su naturaleza se expande al incorporar grandes cantidades de datos no estructurados y semiestructurados, lo que exige el uso de tecnologías especializadas en almacenamiento y procesamiento como Hadoop y sistemas de bases de datos NoSQL.
    d) Además de su complejidad y diversidad, Big Data introduce desafíos significativos en términos de veracidad y variabilidad al requerir métodos avanzados de limpieza y validación de datos para asegurar la integridad del análisis.

11. En la RI en el contexto de Big Data, se puede asegurar que:
    a) El almacenamiento distribuido centraliza todos los datos en un único servidor para facilitar su gestión y mantenimiento.
    b) El almacenamiento distribuido ofrece ventajas significativas en términos de escalabilidad y rendimiento en comparación con el almacenamiento centralizado.
    c) El almacenamiento distribuido reparte los datos en múltiples servidores para mejorar la disponibilidad y la redundancia del sistema.
    d) El almacenamiento distribuido es una técnica obsoleta.

12. Existen varias estrategias consideradas como no recomendables para mejorar el SEO de un sitio web. Dentro de estas estrategias negativas, se encuentran:
    a) Construir enlaces naturales de sitios web con autoridad y relevancia temática.
    b) Evitar el uso de las meta etiquetas y descripciones del sitio web para reflejar el contenido de la página.
    c) Incluir una cantidad excesiva de palabras clave irrelevantes para intentar manipular los rankings de búsqueda.
    d) Utilizar técnicas de encubrimiento para mostrar contenido diferente a lo indexado por los motores de búsqueda.

13. Para contribuir positivamente al posicionamiento orgánico de un sitio web en los motores de búsqueda se puede:
    a) Crear contenido relevante y de alta calidad que satisfaga las necesidades de información de los usuarios.
    b) Mejorar la velocidad de carga del sitio web y asegurar que sea *responsive* y fácil de usar en dispositivos móviles.
    c) Incluir una densidad alta de palabras clave para asegurar que el sitio web aparezca en tantas búsquedas como sea posible.
    d) Obtener enlaces entrantes de otros sitios web de alta autoridad y relevancia temática.

14. Los Web Crawlers se enfrentan a desafíos constantes. Dentro de ellos se encuentran:
    a) La dificultad para generar el contenido dinámico en tiempo real.
    b) La modificación del código y la estructura del sitio web.
    c) La incapacidad para interpretar correctamente el lenguaje de programación utilizado en el desarrollo de los sitios web.
    d) La falta de acceso a la base de datos del servidor web para extraer información actualizada.

15. El algoritmo Hypertext Induced Topic Selection (HITS) intenta buscar nodos especiales. Estos son conocidos como:
    a) Autoridades.
    b) Hubs.
    c) Centrales.
    d) Sensibles.

16. En un sistema de control de tráfico urbano basado en reglas, ¿cuál de las siguientes reglas sería más efectiva para manejar situaciones de congestión en una intersección?
    a) Si hay pocos vehículos en la intersección, reducir el tiempo de los semáforos en verde.
    b) Si hay muchos vehículos en la intersección, aumentar el tiempo de los semáforos en verde.
    c) Si hay un vehículo de emergencia en la intersección, detener todos los demás vehículos.
    d) Si hay muchos peatones cruzando la intersección, reducir el tiempo de los semáforos en rojo.

17. El componente responsable de la gestión de recursos y planificación de tareas en Hadoop es:
    a) MapReduce.
    b) Hadoop Common.
    c) HDFS.
    d) YARN.

18. El algoritmo de PageRank puede describirse como un procedimiento utilizado para:
    a) Determinar la velocidad de carga de una página web en un navegador.
    b) Calcular la relevancia de una página web en función de la cantidad y calidad de los enlaces que apuntan hacia ella.
    c) Prevenir el spam y el contenido no deseado en las páginas web.
    d) Clasificar las páginas web en función de su edad y autoridad.

19. En el modelo de representación del conocimiento basado en herencia se puede asegurar que:
    a) La cantidad de padres que puede tener un nodo no es mayor que 4.
    b) Las conclusiones pueden ser canceladas si el grafo es ambiguo.
    c) Solo se usa en entornos referentes a la biología.
    d) El razonamiento deducido está respaldado por al menos un camino dentro del grafo.

20. La Web 3 se conoce como:
    a) Internet de las cosas.
    b) Web de solo lectura.
    c) Web semántica.
    d) Web de escritura-lectura.

Trabajo de Control
Sistemas de Recuperación de Información
Fecha: 15 de julio de 2024
Temario #57

Nombre: _____ Grupo: _____

Seleccione en cada caso la(s) respuesta(s) correcta(s). Considere también no marcar ninguna opción.

1. En una biblioteca digital se necesita implementar un sistema que permita a los usuarios encontrar libros y artículos científicos de forma rápida y precisa. Los documentos están en diversos formatos, incluyendo PDF, EPUB y HTML. Se requiere seleccionar un algoritmo de indexación adecuado para el sistema, por lo que el programador designado para la implementación debe considerar:
   a) La velocidad de indexación y recuperación de datos.
   b) La capacidad para manejar documentos en diferentes formatos.
   c) La complejidad del algoritmo en términos de implementación y mantenimiento.
   d) La capacidad del algoritmo para procesar imágenes incrustadas o referenciadas en los ficheros.

2. Analizar una red permite:
   a) Detectar posibles tendencias antes de que se conviertan en tendencia.
   b) Obtener predicciones exactas de eventos futuros en mercados financieros.
   c) Evaluar la calidad del contenido de un sitio web.
   d) Encontrar nodos "sensibles" o críticos para la red.

3. En el diseño de un videojuego de roles (RPG) en el que los personajes tienen atributos como salud, fuerza y velocidad, ¿cuál de las siguientes opciones representa mejor una implementación de la representación del conocimiento orientado a objetos?
   a) Cada personaje se representa como una lista de cadenas de texto que describen sus características físicas y habilidades.
   b) Cada personaje se representa como una función que calcula sus atributos en función de su nivel y experiencia.
   c) Cada personaje se representa como una matriz de números que almacena sus valores de atributos.
   d) Cada personaje se representa como un objeto con propiedades como salud, fuerza y velocidad, y métodos para modificar y consultar estos valores.

4. Una plataforma de comercio electrónico desea mejorar la experiencia del usuario al permitir una navegación más personalizada y contextualizada. Actualmente los usuarios tienen dificultades para encontrar productos relevantes debido a la gran cantidad de opciones disponibles. La empresa está interesada en implementar características de la Web 2.5 y la Web Semántica para abordar este problema. ¿Qué características podrían ayudar para ofrecer una navegación más personalizada y contextualizada?
   a) La implementación de ontologías y metadatos para enriquecer la descripción de productos y mejorar la precisión de las recomendaciones.
   b) La optimización de la velocidad de carga del sitio web para mejorar la experiencia del usuario y reducir el abandono del carrito de compra.
   c) La integración de redes sociales para permitir la recomendación de productos basada en las preferencias de amigos y contactos.
   d) La incorporación de los productos en tendencia en el mercado.

5. El uso de Hadoop y MapReduce en el contexto de la RI tiene como objetivo:
   a) Automatizar el mantenimiento de sistemas de bases de datos.
   b) Facilitar el análisis en tiempo real de datos de redes sociales.
   c) Mejorar la eficiencia energética en centros de datos.
   d) Procesar y analizar grandes conjuntos de datos para la RI.

6. En el contexto del procesamiento de grandes conjuntos de datos, se puede asegurar que:
   a) MapReduce es un enfoque para procesar datos de forma secuencial en un solo servidor para evitar problemas de concurrencia.
   b) MapReduce divide una tarea en múltiples pasos de map y reduce que se ejecutan de forma secuencial en diferentes servidores para mejorar el rendimiento y la escalabilidad.
   c) MapReduce no es adecuado para el procesamiento de datos no estructurados.
   d) MapReduce solo puede manejar pequeñas cantidades de datos y no escala bien a grandes conjuntos de datos.

7. La integración de Hadoop y MapReduce en la RI trae como ventaja:

a) La eliminación de la necesidad de sistemas de bases de datos.

b) La posibilidad de la extracción de información relevante y la generación de resultados significativos de grandes conjuntos de datos.

c) La garantía de la privacidad absoluta de los datos procesados.

d) La reducción de los costos operativos a cero.

8. En el contexto de la representación del conocimiento basada en herencia, ¿qué caracteriza a la herencia cancelable?

a) La herencia es el resultado del razonamiento no transitivo.

b) Las conclusiones no están determinadas y dependen del nodo de interés.

c) Las propiedades heredadas siempre se mantienen y no pueden anularse.

d) No existe ambigüedad en las conclusiones obtenidas.

9. Los algoritmos para detectar comunidades en una red intentan:

a) Buscar conjuntos donde cada nodo de un mismo conjunto tenga características similares al resto de los nodos del conjunto.

b) Seleccionar aleatoriamente nodos de alto grado y sus vecinos.

c) Buscar subgrafos tal que no incluyan nodos cuya ausencia desconecte al subgrafo.

d) Encontrar grupos donde los nodos pertenecientes a los mismos grupos son cercanos bajo cierta métrica y lejanos con respecto a los nodos de otros grupos.

10. El algoritmo de PageRank puede describirse como un procedimiento utilizado para:

a) Determinar la velocidad de carga de una página web en un navegador.

b) Calcular la relevancia de una página web en función de la cantidad y calidad de los enlaces que apuntan hacia ella.

c) Prevenir el spam y el contenido no deseado en las páginas web.

d) Clasificar las páginas web en función de su edad y autoridad.

11. ¿Cuál de las siguientes opciones describe mejor la diferencia clave entre Web Crawling y Web Scraping?

a) Web Crawling se centra en la exploración y recopilación de enlaces de múltiples sitios web, mientras que Web Scraping se enfoca en la extracción específica de datos de páginas web individuales.

b) Web Crawling se realiza utilizando herramientas de automatización como Selenium WebDriver, mientras que Web Scraping se lleva a cabo mediante el análisis de HTML y CSS.

c) Web Scraping implica el análisis de la estructura y el contenido de las páginas web para extraer datos, mientras que Web Crawling se refiere a la descarga y almacenamiento de páginas web completas.

d) Web Scraping es más eficaz para rastrear e indexar contenido web para motores de búsqueda, mientras que Web Crawling se utiliza principalmente para la extracción de datos en proyectos de investigación.

12. El algoritmo Hypertext Induced Topic Selection (HITS) intenta buscar nodos especiales. Estos son conocidos como:

a) Autoridades.

b) Hubs.

c) Centrales.

d) Sensibles.

13. Dentro del análisis de redes, la centralidad de grado mide la importancia de un nodo basándose en:

a) El grado del nodo.

b) La cantidad de veces que aparece el nodo en el camino mínimo entre cualquier par de nodos.

c) La cantidad de vecinos del nodo.

d) El número de aristas que posee el nodo.

14. La computación evolutiva:

a) No es aplicable en la RI debido a la complejidad de los algoritmos evolutivos.

b) Solo puede manejar conjuntos de datos pequeños y no es escalable a grandes volúmenes de datos.

c) Utiliza algoritmos para buscar soluciones óptimas en grandes espacios de búsqueda, lo que la hace adecuada para problemas de optimización en la RI.

d) Es útil solo para problemas de clasificación de documentos y no para otras tareas de RI en general.

15. El concepto de *rank sink* en el algoritmo de PageRank representa páginas web:

a) Con distribuciones uniformes de la puntuación de PageRank.

b) Que tienen una puntuación de PageRank más alta que otras debido a la manipulación de enlaces entrantes y salientes.

c) Con una baja calidad de contenido y una cantidad insuficiente de enlaces salientes, lo que las hace menos relevantes en los resultados de búsqueda.

d) Con un alto número de enlaces salientes que no reciben enlaces entrantes, lo que puede afectar negativamente su puntuación de PageRank.

16. ¿Qué es un “Uniform Resource Locator (URL) Frontier” en el contexto de Web Crawling?

a) Un protocolo que define cómo se deben formatear las URLs para el crawling.

b) Una lista prioritaria de URLs que aún no han sido visitadas por el crawler.

c) Una técnica para filtrar URLs irrelevantes y mejorar la eficiencia del crawling.

d) Una base de datos que almacena URLs únicas identificadas como recursos en la web.

17. Con respecto a WordNet y su contribución a las ontologías y la representación del conocimiento en los sistemas de RI, puede afirmarse que:

a) En WordNet, una palabra está asociada a un conjunto de sinónimos (synsets), siendo estas palabras intercambiables en un contexto.

b) La integración de WordNet en sistemas de RI permite la expansión de consultas y la mejora de la precisión de los resultados al entender mejor el significado de los términos de búsqueda a través de su contexto semántico y las relaciones entre palabras.

c) WordNet diferencia claramente entre los significados de palabras según su uso en diferentes contextos, lo que permite aplicaciones avanzadas en desambiguación semántica más allá de los sistemas de recomendación y RI.

d) Una limitación de WordNet en la representación del conocimiento es su enfoque en el idioma inglés, lo que plantea desafíos en la aplicación global y la interoperabilidad con sistemas de información multilingües.

18. Al "relajar" el concepto de clique en la detección de comunidades se intenta solucionar:

a) La imposibilidad de conectar nodos distantes.

b) La necesidad de datos externos para analizar la red.

c) La uniformidad de los nodos en términos de grado.

d) El solapamiento y la complejidad computacional.

19. El posicionamiento de un sitio web en los motores de búsqueda puede ser afectado por:

a) Utilizar técnicas de encubrimiento para mostrar contenido diferente a lo indexado por los motores de búsqueda.

b) Contenido sin valor en el sitio web.

c) Mantener una estructura de URL clara y coherente.

d) Obtener enlaces de sitios web irrelevantes y de baja calidad.

20. En un sistema donde el conocimiento está definido a partir de reglas se puede asegurar que:

a) El nivel de especificidad de las reglas está limitado.

b) La representación de conocimiento está basada en la lógica proposicional.

c) El razonamiento se activa en cadena hacia delante.

d) El orden en que se definen las reglas no altera el razonamiento del sistema.

Trabajo de Control
Sistemas de Recuperación de Información
Fecha: 15 de julio de 2024
Temario #58

Nombre: _____ Grupo: _____

Seleccione en cada caso la(s) respuesta(s) correcta(s). Considere también no marcar ninguna opción.

1. Se puede afirmar que:
   a) El término "Big Data" se refiere exclusivamente al volumen de datos que una organización maneja, sin tener en cuenta la velocidad, la variedad y la veracidad de los datos.
   b) Los SRI pueden beneficiarse de MapReduce para mejorar la RI relevante.
   c) MapReduce es un modelo de procesamiento distribuido utilizado para trabajar con grandes volúmenes de datos.
   d) Uno de los desafíos en el procesamiento de Big Data es la capacidad de gestionar y analizar datos provenientes de diversas fuentes y en diferentes formatos de manera eficiente.

2. En una red de transporte donde cada nodo es una parada de autobús y las aristas representan si existe un carro que pasa por ambos sitios, ¿qué puede mejorar el sistema de transporte?
   a) No considerar la centralidad de intermediación de las estaciones de transporte público al planificar rutas y horarios, ya que no tiene impacto en las conexiones entre las paradas.
   b) Utilizar el grafo inducido de los nodos con mayor valor en la centralidad de grado para aplicar la centralidad de intermediación con el propósito de reforzar las paradas con mayor tráfico.
   c) Utilizar la centralidad de intermediación para identificar las paradas de transferencia clave y establecer nuevas conexiones entre diferentes líneas de transporte público.
   d) No implementar sistemas de información en tiempo real para los usuarios, ya que pueden aumentar la carga de los trabajadores.

3. Sobre WordNet y su aplicación en el procesamiento del lenguaje natural, se puede afirmar que:
   a) El diseño de WordNet facilita su integración en aplicaciones multilingües de NLP, aunque su desarrollo original se centró en el inglés.
   b) Aunque WordNet es una herramienta valiosa en el NLP, su estructura no incluye información sobre la frecuencia de uso de las palabras en el lenguaje natural.
   c) Los synsets facilitan la identificación de relaciones semánticas entre palabras, como la hiperonimia y la meronimia, enriqueciendo tareas de NLP.
   d) WordNet proporciona una base para la desambiguación semántica al agrupar palabras con significados similares en synsets.

4. En el contexto de la representación del conocimiento basada en herencia, ¿qué caracteriza a la herencia cancelable?
   a) La herencia es el resultado del razonamiento no transitivo.
   b) Las conclusiones no están determinadas y dependen del nodo de interés.
   c) Las propiedades heredadas siempre se mantienen y no pueden anularse.
   d) No existe ambigüedad en las conclusiones obtenidas.

5. Se tiene una red de ingredientes donde cada uno representa un nodo y las aristas simbolizan que los ingredientes forman parte de una misma receta. Se busca mejorar la experiencia culinaria mediante la elaboración de combinaciones de ingredientes más interesantes y creativas, para lo cual se debe:
   a) No considerar la centralidad de grado de los ingredientes, ya que todas las combinaciones de ingredientes son igualmente válidas.
   b) Utilizar la centralidad de grado para identificar los ingredientes menos conectados en la red y tomarlos en cuenta para su inclusión en las combinaciones.
   c) No tener en cuenta la centralidad de grado de los ingredientes para no darle mayor importancia a los ingredientes más comunes.
   d) Utilizar la centralidad de grado para identificar los ingredientes más populares en la red y crear combinaciones que incluyan una variedad de ingredientes menos comunes.

6. El propósito de la política de revisitado en los Web Crawlers es:
   a) Determinar la frecuencia con la que los crawlers deben volver a visitar una URL específica para mantener la información actualizada en el índice de búsqueda.

b) Establecer reglas sobre el tiempo máximo que los crawlers pueden pasar en un sitio web durante cada visita para evitar sobrecargar los servidores.

c) Limitar el acceso de los crawlers a ciertos servidores luego de visitar las páginas alojadas en estos.

d) Definir la prioridad de rastreo de las URLs, determinando el orden en que los crawlers visitan y procesan cada página web.

7. La centralidad de intermediación de un nodo indica:

a) La cantidad de nodos vecinos directos.

b) La resistencia del nodo a fallos.

c) El número total de conexiones entrantes y salientes.

d) La frecuencia con la que un nodo actúa como puente en el camino más corto entre otros dos nodos.

8. Un investigador necesita recopilar datos de múltiples sitios web para un estudio académico, pero se enfrenta a varios desafíos al realizar el proceso de Web Scraping de manera ética y legal. ¿Cuál de las siguientes opciones describe mejor uno de los desafíos asociados al proceso de Web Scraping?

a) La necesidad de comprender la estructura del sitio web y su código HTML para extraer los datos correctamente.

b) La disponibilidad limitada de datos en línea que se pueden extraer utilizando técnicas de Web Scraping.

c) La dificultad para encontrar herramientas de Web Scraping gratuitas y fiables.

d) La necesidad de estar montado sobre un scrawler que cumpla con todas las políticas.

9. La web actual se enfrenta a problemas como:

a) Presencia de grandes volúmenes de datos estructurados.

b) Presencia de una alta calidad en los datos.

c) Presencia de datos volátiles y distribuidos.

d) Heterogeneidad en los datos.

10. La Web 3 se conoce como:

a) Internet de las cosas.

b) Web de solo lectura.

c) Web semántica.

d) Web de escritura-lectura.

11. ¿Cuál de las siguientes opciones describe mejor la diferencia clave entre Web Crawling y Web Scraping?

a) Web Crawling se centra en la exploración y recopilación de enlaces de múltiples sitios web, mientras que Web Scraping se enfoca en la extracción específica de datos de páginas web individuales.

b) Web Crawling se realiza utilizando herramientas de automatización como Selenium WebDriver, mientras que Web Scraping se lleva a cabo mediante el análisis de HTML y CSS.

c) Web Scraping implica el análisis de la estructura y el contenido de las páginas web para extraer datos, mientras que Web Crawling se refiere a la descarga y almacenamiento de páginas web completas.

d) Web Scraping es más eficaz para rastrear e indexar contenido web para motores de búsqueda, mientras que Web Crawling se utiliza principalmente para la extracción de datos en proyectos de investigación.

12. Sobre el algoritmo de PageRank, visto en clase, se puede afirmar que:

a) Evalúa la importancia de un sitio web en función de la calidad y cantidad de enlaces entrantes que recibe de otros sitios web.

b) Solo tiene en cuenta el contenido en un sitio web para determinar su relevancia en los resultados de búsqueda.

c) Asigna una puntuación alta a los sitios web que tienen un gran número de enlaces entrantes sin tener en cuenta la calidad de esos enlaces.

d) Asigna una puntuación baja a los sitios web que contienen muchos enlaces salientes, ya que indica una falta de relevancia.

13. Con respecto a WordNet y su contribución a las ontologías y la representación del conocimiento en los sistemas de RI, puede afirmarse que:

a) En WordNet, una palabra está asociada a un conjunto de sinónimos (synsets), siendo estas palabras intercambiables en un contexto.

b) La integración de WordNet en sistemas de RI permite la expansión de consultas y la mejora de la precisión de los resultados al entender mejor el significado de los términos de búsqueda a través de su contexto semántico y las relaciones entre palabras.

c) WordNet diferencia claramente entre los significados de palabras según su uso en diferentes contextos, lo que permite aplicaciones avanzadas en desambiguación semántica más allá de los sistemas de recomendación y RI.

d) Una limitación de WordNet en la representación del conocimiento es su enfoque en el idioma inglés, lo que plantea desafíos en la aplicación global y la interoperabilidad con sistemas de información multilingües.

14. El componente responsable de la gestión de recursos y planificación de tareas en Hadoop es:

a) MapReduce.

b) Hadoop Common.

c) HDFS.

d) YARN.

15. El posicionamiento de un sitio web en los motores de búsqueda puede ser afectado por:
    a) Utilizar técnicas de encubrimiento para mostrar contenido diferente a lo indexado por los motores de búsqueda.
    b) Contenido sin valor en el sitio web.
    c) Mantener una estructura de URL clara y coherente.
    d) Obtener enlaces de sitios web irrelevantes y de baja calidad.

16. En la RI en el contexto de Big Data, se puede asegurar que:
    a) El almacenamiento distribuido centraliza todos los datos en un único servidor para facilitar su gestión y mantenimiento.
    b) El almacenamiento distribuido ofrece ventajas significativas en términos de escalabilidad y rendimiento en comparación con el almacenamiento centralizado.
    c) El almacenamiento distribuido reparte los datos en múltiples servidores para mejorar la disponibilidad y la redundancia del sistema.
    d) El almacenamiento distribuido es una técnica obsoleta.

17. Para contribuir positivamente al posicionamiento orgánico de un sitio web en los motores de búsqueda se puede:
    a) Crear contenido relevante y de alta calidad que satisfaga las necesidades de información de los usuarios.
    b) Mejorar la velocidad de carga del sitio web y asegurar que sea *responsive* y fácil de usar en dispositivos móviles.
    c) Incluir una densidad alta de palabras clave para asegurar que el sitio web aparezca en tantas búsquedas como sea posible.
    d) Obtener enlaces entrantes de otros sitios web de alta autoridad y relevancia temática.

18. Se puede afirmar que:
    a) No es posible identificar subgrupos dentro de una red utilizando análisis de redes.
    b) La cantidad de conexiones de un nodo siempre indica su influencia en la red.
    c) El tamaño de una red es siempre indicativo de su efectividad en la transmisión de información.
    d) Todas las relaciones en una red tienen la misma importancia para el análisis.

19. Para transformar el contenido no estructurado de las páginas web en datos estructurados, el Web Scraping:
    a) Extrae información basada en patrones de HTML/CSS.
    b) Analiza los protocolos de red.
    c) Interpreta el código JavaScript en tiempo real.
    d) Convierte de forma automática imágenes a texto.

20. La transición de la Web 1.0 a la Web 2.0 se caracterizó principalmente por:
    a) El aumento en la velocidad de conexión a internet, que permitió una mejor calidad de las páginas web.
    b) La disminución de la importancia de los motores de búsqueda en la navegación web.
    c) El cambio de páginas web estáticas a dinámicas, permitiendo la interacción del usuario y la generación de contenido.
    d) La reducción en el uso de HTML y CSS en el desarrollo de sitios web.

Trabajo de Control
Sistemas de Recuperación de Información
Fecha: 15 de julio de 2024
Temario #59

Nombre: _____ Grupo: _____

Seleccione en cada caso la(s) respuesta(s) correcta(s). Considere también no marcar ninguna opción.

1. Considerando las prácticas éticas y legales en el Web Scraping, se puede asegurar que:
   a) Web Scraping requiere considerar las políticas de `robots.txt` del sitio web objetivo.
   b) Es importante revisar y respetar los términos de servicio del sitio web, así como las leyes aplicables de protección de datos y derechos de autor, antes de realizar Web Scraping.
   c) Web Scraping sobre datos personales sin consentimiento es generalmente aceptado si los datos se utilizan con fines de investigación.
   d) La extracción de datos mediante Web Scraping siempre es legal, independientemente de las leyes locales sobre derechos de autor y privacidad de datos definidos en los sitios web.

2. La computación evolutiva:
   a) No es aplicable en la RI debido a la complejidad de los algoritmos evolutivos.
   b) Solo puede manejar conjuntos de datos pequeños y no es escalable a grandes volúmenes de datos.
   c) Utiliza algoritmos para buscar soluciones óptimas en grandes espacios de búsqueda, lo que la hace adecuada para problemas de optimización en la RI.
   d) Es útil solo para problemas de clasificación de documentos y no para otras tareas de RI en general.

3. En una biblioteca digital se necesita implementar un sistema que permita a los usuarios encontrar libros y artículos científicos de forma rápida y precisa. Los documentos están en diversos formatos, incluyendo PDF, EPUB y HTML. Se requiere seleccionar un algoritmo de indexación adecuado para el sistema, por lo que el programador designado para la implementación debe considerar:
   a) La velocidad de indexación y recuperación de datos.
   b) La capacidad para manejar documentos en diferentes formatos.
   c) La complejidad del algoritmo en términos de implementación y mantenimiento.
   d) La capacidad del algoritmo para procesar imágenes incrustadas o referenciadas en los ficheros.

4. En la RI en el contexto de Big Data, se puede asegurar que:
   a) La indexación distribuida divide los datos en múltiples fragmentos que se almacenan en varios nodos para permitir búsquedas paralelas y mejorar la escalabilidad.
   b) La indexación distribuida no ofrece ventajas en términos de rendimiento y escalabilidad en comparación con la indexación centralizada.
   c) La indexación distribuida es una técnica obsoleta.
   d) La indexación distribuida almacena todos los datos en un solo servidor para facilitar su acceso y búsqueda.

5. El algoritmo de PageRank converge si:
   a) La norma de la diferencia entre los vectores es menor a un umbral predefinido.
   b) Finaliza la ejecución cuando el número de iteraciones excede un máximo de iteraciones previamente definido.
   c) Se define un factor de normalización en la fórmula de la función.
   d) El algoritmo no se implementa de forma iterativa.

6. La política de ordenación de URLs en los Web Crawlers tiene como aspecto fundamental:
   a) Limitar el acceso de los crawlers a ciertas secciones de un sitio web, evitando el rastreo de URLs consideradas menos importantes o sensibles.
   b) Definir la estructura de la URL de destino, asegurando que estén ordenadas alfabéticamente para facilitar la navegación y la indexación.
   c) Establecer la prioridad de rastreo de las URLs, determinando el orden en que los crawlers visitan y procesan cada página web.
   d) Determinar la forma en que los crawlers asignan un valor de relevancia a cada URL para clasificarlas en el índice de búsqueda.

7. ¿Cuál de las siguientes opciones describe mejor la diferencia clave entre Web Crawling y Web Scraping?
   a) Web Crawling se centra en la exploración y recopilación de enlaces de múltiples sitios web, mientras que Web Scraping se enfoca en la extracción específica de datos de páginas web individuales.

b) Web Crawling se realiza utilizando herramientas de automatización como Selenium WebDriver, mientras que Web Scraping se lleva a cabo mediante el análisis de HTML y CSS.

c) Web Scraping implica el análisis de la estructura y el contenido de las páginas web para extraer datos, mientras que Web Crawling se refiere a la descarga y almacenamiento de páginas web completas.

d) Web Scraping es más eficaz para rastrear e indexar contenido web para motores de búsqueda, mientras que Web Crawling se utiliza principalmente para la extracción de datos en proyectos de investigación.

8. Dentro del análisis de redes, la centralidad de grado mide la importancia de un nodo basándose en:
   a) El grado del nodo.
   b) La cantidad de veces que aparece el nodo en el camino mínimo entre cualquier par de nodos.
   c) La cantidad de vecinos del nodo.
   d) El número de aristas que posee el nodo.

9. En el modelo de representación del conocimiento basado en herencia se puede asegurar que:
   a) La cantidad de padres que puede tener un nodo no es mayor que 4.
   b) Las conclusiones pueden ser canceladas si el grafo es ambiguo.
   c) Solo se usa en entornos referentes a la biología.
   d) El razonamiento deducido está respaldado por al menos un camino dentro del grafo.

10. Al realizar la optimización de contenido para SEO debe considerarse:
    a) Seleccionar las palabras clave al azar.
    b) Utilizar etiquetas de título y meta descripciones únicas y relevantes para cada página.
    c) Incluir palabras clave de manera excesiva en el contenido para mejorar el posicionamiento.
    d) Crear contenido valioso y original que satisfaga las necesidades de los usuarios.

11. En un sistema donde el conocimiento está orientado a objetos se puede asegurar que:
    a) Se enfatiza la atención a la información de la cual se extrajo el conocimiento.
    b) Los marcos y las bandas son las estructuras utilizadas para la representación del modelo.
    c) No es posible definir especializaciones de los objetos de la vida real dentro del sistema.
    d) Las funciones de agregación dificultan poder establecer relaciones entre los objetos.

12. Para la RI, el análisis de las redes puede:
    a) Ayudar a identificar grupos de interés.
    b) Indicar la importancia de una entidad en la transmisión de la información.
    c) Ayudar a comprender la conectividad y la accesibilidad entre las entidades.
    d) Revelar patrones de influencia dentro de una comunidad.

13. Se tiene una red de ingredientes donde cada uno representa un nodo y las aristas simbolizan que los ingredientes forman parte de una misma receta. Se busca mejorar la experiencia culinaria mediante la elaboración de combinaciones de ingredientes más interesantes y creativas, para lo cual se debe:
    a) No considerar la centralidad de grado de los ingredientes, ya que todas las combinaciones de ingredientes son igualmente válidas.
    b) Utilizar la centralidad de grado para identificar los ingredientes menos conectados en la red y tomarlos en cuenta para su inclusión en las combinaciones.
    c) No tener en cuenta la centralidad de grado de los ingredientes para no darle mayor importancia a los ingredientes más comunes.
    d) Utilizar la centralidad de grado para identificar los ingredientes más populares en la red y crear combinaciones que incluyan una variedad de ingredientes menos comunes.

14. El conjunto frontera de URLs en un Web Crawler:
    a) Delimita los sitios web que visitará en futuras iteraciones del proceso.
    b) Es similar a un conjunto de URLs que esperan ser visitadas.
    c) Almacena hipervínculos que pertenecen al mismo dominio del conjunto semilla de URLs con que inició el crawler.
    d) Indica las secciones que no pueden ser visitadas de cada sitio web.

15. Una buena práctica de SEO para mejorar el posicionamiento de un sitio web en los motores de búsqueda es:
    a) Obtener enlaces de otros sitios web relevantes y de calidad que apunten al sitio.
    b) Copiar contenido directamente de otros sitios web populares para aumentar la cantidad de páginas indexadas.
    c) Utilizar técnicas de encubrimiento para mostrar contenido diferente a lo indexado por los motores de búsqueda.
    d) Llenar el contenido con palabras clave irrelevantes para aumentar la densidad de palabras clave.

16. La transición de la Web 1.0 a la Web 2.0 se caracterizó principalmente por:

a) El aumento en la velocidad de conexión a internet, que permitió una mejor calidad de las páginas web.

b) La disminución de la importancia de los motores de búsqueda en la navegación web.

c) El cambio de páginas web estáticas a dinámicas, permitiendo la interacción del usuario y la generación de contenido.

d) La reducción en el uso de HTML y CSS en el desarrollo de sitios web.

17. En un SRI la indexación:

a) Consiste en asociar un identificador único a cada dato almacenado en el sistema.

b) Mejora la experiencia del usuario.

c) Permite la organización y la categorización de la información.

d) Facilita la RI relevante.

18. El concepto de *rank sink* en el algoritmo de PageRank representa páginas web:

a) Con distribuciones uniformes de la puntuación de PageRank.

b) Que tienen una puntuación de PageRank más alta que otras debido a la manipulación de enlaces entrantes y salientes.

c) Con una baja calidad de contenido y una cantidad insuficiente de enlaces salientes, lo que las hace menos relevantes en los resultados de búsqueda.

d) Con un alto número de enlaces salientes que no reciben enlaces entrantes, lo que puede afectar negativamente su puntuación de PageRank.

19. En el contexto de la RI en redes sociales se puede afirmar que:

a) El análisis de redes sociales se centra exclusivamente en la cantidad de seguidores que tiene un usuario en particular para determinar su influencia en la red.

b) El análisis de centralidad de intermediación se utiliza para identificar usuarios que son importantes en una red social debido a su posición como "puentes" entre diferentes grupos de usuarios.

c) El análisis de sentimientos se utiliza para determinar la popularidad de una publicación en redes sociales sin tener en cuenta la opinión de los usuarios.

d) El análisis de redes sociales no es útil para comprender la difusión de información en una red social específica, para ello se utiliza la medida de centralidad de vector propio.

20. Dentro del ecosistema de Hadoop, el HDFS se caracteriza por:

a) El modelo de acceso y de escritura de datos en tiempo real.

b) La tolerancia a fallos mediante la replicación de datos.

c) El almacenamiento exclusivo para archivos de texto.

d) La capacidad ilimitada de almacenamiento.

Trabajo de Control
Sistemas de Recuperación de Información
Fecha: 15 de julio de 2024
Temario #60

Nombre: _____ Grupo: _____

Seleccione en cada caso la(s) respuesta(s) correcta(s). Considere también no marcar ninguna opción.

1. ¿Qué estrategia utilizan los Web Crawlers para asegurar un rastreo eficiente y respetuoso de los recursos de los sitios web?
   a) Ignorar completamente el archivo `robots.txt` de los sitios web.
   b) Visitar y rastrear todos los enlaces de una página web simultáneamente.
   c) Extraer únicamente contenido multimedia para reducir la carga en los servidores web.
   d) Seguir las directrices del archivo `robots.txt` y aplicar un retraso entre las solicitudes.

2. Referente al Web Crawler puede afirmarse que:
   a) Los hipervínculos encontrados en cada sitio web que no pertenecen al dominio donde fueron encontrados se desechan, puesto que no expande el conjunto de URLs sin visitar.
   b) Las páginas visitadas no se procesan nunca más.
   c) No necesita de un conjunto inicial de URLs para recorrer la Web.
   d) No tiene como objetivo indexar y recopilar información de diferentes sitios web.

3. En el contexto de la RI en redes sociales se puede afirmar que:
   a) El análisis de redes sociales se centra exclusivamente en la cantidad de seguidores que tiene un usuario en particular para determinar su influencia en la red.
   b) El análisis de centralidad de intermediación se utiliza para identificar usuarios que son importantes en una red social debido a su posición como "puentes" entre diferentes grupos de usuarios.
   c) El análisis de sentimientos se utiliza para determinar la popularidad de una publicación en redes sociales sin tener en cuenta la opinión de los usuarios.
   d) El análisis de redes sociales no es útil para comprender la difusión de información en una red social específica, para ello se utiliza la medida de centralidad de vector propio.

4. Si se tiene el conjunto de páginas interconectadas $\{A \to B, C; B \to C; C \to A\}$, entonces la página con valor más alto de PageRank es:
   a) A y B.
   b) A.
   c) C.
   d) B.

5. El conjunto frontera de URLs en un Web Crawler:
   a) Delimita los sitios web que visitará en futuras iteraciones del proceso.
   b) Es similar a un conjunto de URLs que esperan ser visitadas.
   c) Almacena hipervínculos que pertenecen al mismo dominio del conjunto semilla de URLs con que inició el crawler.
   d) Indica las secciones que no pueden ser visitadas de cada sitio web.

6. La afirmación que mejor describe la política de amabilidad en los Web Crawlers es:
   a) Los crawlers se diseñan para acceder a sitios web sin restricciones y extraer datos de manera agresiva para su indexación.
   b) La política de amabilidad de los Web Crawlers dicta que los crawlers deben priorizar ciertos tipos de contenido sobre otros, ignorando completamente ciertas páginas web.
   c) La política de amabilidad establece pautas y reglas sobre cómo los crawlers deben interactuar con los sitios web para minimizar la carga del servidor y respetar las directivas de los administradores del sitio.
   d) Los Web Crawlers son libres de recopilar datos de cualquier sitio web sin restricciones, independientemente de la cantidad de tráfico que generen.

7. Analizar una red permite:
   a) Detectar posibles tendencias antes de que se conviertan en tendencia.
   b) Obtener predicciones exactas de eventos futuros en mercados financieros.
   c) Evaluar la calidad del contenido de un sitio web.
   d) Encontrar nodos "sensibles" o críticos para la red.

8. Sobre el algoritmo de PageRank, visto en clase, se puede afirmar que:
   a) Evalúa la importancia de un sitio web en función de la calidad y cantidad de enlaces entrantes que recibe de otros sitios web.

b) Solo tiene en cuenta el contenido en un sitio web para determinar su relevancia en los resultados de búsqueda.

c) Asigna una puntuación alta a los sitios web que tienen un gran número de enlaces entrantes sin tener en cuenta la calidad de esos enlaces.

d) Asigna una puntuación baja a los sitios web que contienen muchos enlaces salientes, ya que indica una falta de relevancia.

9. En el algoritmo de Indexación Basada en Clasificación Bloqueada (BSBI), ¿cuál es el paso final para crear un índice invertido para la colección completa de los datos?

a) Eliminar los términos duplicados de los índices.

b) Indexar cada bloque de forma independiente.

c) Fusionar los índices invertidos de cada bloque.

d) Dividir la colección de datos en bloques de tamaño fijo.

10. Los Web Crawlers se enfrentan a desafíos constantes. Dentro de ellos se encuentran:

a) La dificultad para generar el contenido dinámico en tiempo real.

b) La modificación del código y la estructura del sitio web.

c) La incapacidad para interpretar correctamente el lenguaje de programación utilizado en el desarrollo de los sitios web.

d) La falta de acceso a la base de datos del servidor web para extraer información actualizada.

11. El uso de Hadoop y MapReduce en el contexto de la RI tiene como objetivo:

a) Automatizar el mantenimiento de sistemas de bases de datos.

b) Facilitar el análisis en tiempo real de datos de redes sociales.

c) Mejorar la eficiencia energética en centros de datos.

d) Procesar y analizar grandes conjuntos de datos para la RI.

12. En la RI en el contexto de Big Data, se puede asegurar que:

a) La indexación distribuida divide los datos en múltiples fragmentos que se almacenan en varios nodos para permitir búsquedas paralelas y mejorar la escalabilidad.

b) La indexación distribuida no ofrece ventajas en términos de rendimiento y escalabilidad en comparación con la indexación centralizada.

c) La indexación distribuida es una técnica obsoleta.

d) La indexación distribuida almacena todos los datos en un solo servidor para facilitar su acceso y búsqueda.

13. Se tiene un grafo donde cada nodo es un personaje de cierto libro de cuentos y la existencia de las aristas está definida si dos personajes aparecen en el mismo cuento. Se puede asegurar que:

a) La centralidad de intermediación identifica a los personajes que actúan como conectores entre personajes de cuentos distintos.

b) La centralidad de grado es útil para identificar los personajes con más conexiones dentro de la red, lo que puede indicar su importancia en el libro.

c) La centralidad de cercanía ofrece una relación entre la cantidad de vecinos de un nodo con respecto a la longitud máxima de un camino dentro del grafo partiendo del nodo en cuestión.

d) La centralidad de cercanía indica el grado de conexión de cada personaje con el resto de los personajes de los cuentos del libro.

14. El algoritmo de PageRank puede describirse como un procedimiento utilizado para:

a) Determinar la velocidad de carga de una página web en un navegador.

b) Calcular la relevancia de una página web en función de la cantidad y calidad de los enlaces que apuntan hacia ella.

c) Prevenir el spam y el contenido no deseado en las páginas web.

d) Clasificar las páginas web en función de su edad y autoridad.

15. En un sistema donde el conocimiento está definido a partir de reglas se puede asegurar que:

a) El nivel de especificidad de las reglas está limitado.

b) La representación de conocimiento está basada en la lógica proposicional.

c) El razonamiento se activa en cadena hacia delante.

d) El orden en que se definen las reglas no altera el razonamiento del sistema.

16. En el diseño de un videojuego de roles (RPG) en el que los personajes tienen atributos como salud, fuerza y velocidad, ¿cuál de las siguientes opciones representa mejor una implementación de la representación del conocimiento orientado a objetos?

a) Cada personaje se representa como una lista de cadenas de texto que describen sus características físicas y habilidades.

b) Cada personaje se representa como una función que calcula sus atributos en función de su nivel y experiencia.

c) Cada personaje se representa como una matriz de números que almacena sus valores de atributos.

d) Cada personaje se representa como un objeto con propiedades como salud, fuerza y velocidad, y métodos para modificar y consultar estos valores.

17. Para contribuir positivamente al posicionamiento orgánico de un sitio web en los motores de búsqueda se puede:

a) Crear contenido relevante y de alta calidad que satisfaga las necesidades de información de los usuarios.

b) Mejorar la velocidad de carga del sitio web y asegurar que sea *responsive* y fácil de usar en dispositivos móviles.

c) Incluir una densidad alta de palabras clave para asegurar que el sitio web aparezca en tantas búsquedas como sea posible.

d) Obtener enlaces entrantes de otros sitios web de alta autoridad y relevancia temática.

18. Un sistema cuenta con la siguiente información:

- Juan pertenece a la marina.
- Juan es capellán (sacerdote encargado de una tarea específica fuera de la parroquia).
- Los infantes de la marina suelen ser bebedores de cerveza.
- Un capellán no suele ser bebedor de cerveza.
- Un bebedor de cerveza suele tener sobrepeso.
- Por lo general, un infante de la marina está en buena condición física.

Si se representa la información en una red de herencia se puede concluir que:

a) No se puede asegurar que Juan sea capellán.

b) No se puede asegurar que Juan esté en buena condición física.

c) Juan tiene sobrepeso.

d) Existen dos razonamientos cancelables.

19. Una empresa de comercio electrónico necesita mejorar su motor de búsqueda para proporcionar resultados más relevantes a sus usuarios. Actualmente, los resultados de la búsqueda no son precisos y los usuarios a menudo encuentran dificultades para encontrar productos específicos. La empresa está considerando implementar una indexación por conceptos para mejorar la relevancia de los resultados de búsqueda. ¿Qué beneficios podría provocar este cambio?

a) Permite adaptarse fácilmente a cambios en el vocabulario y la terminología utilizada en las descripciones de los productos.

b) Ayuda a identificar automáticamente relaciones entre productos, lo que puede mejorar las recomendaciones personalizadas a los usuarios.

c) Facilita la visualización de productos al agruparlos por categorías o características comunes.

d) Mejora la precisión en la búsqueda de productos relacionados, incluso cuando no coinciden exactamente con los términos de búsqueda del usuario.

20. Considerando las dimensiones y desafíos inherentes a Big Data puede afirmarse, tomando en cuenta las características clave y las implicaciones para su procesamiento y análisis, que:

a) Big Data se caracteriza principalmente por su pequeño volumen y uniformidad, permitiendo un procesamiento eficiente con mínimas adaptaciones de las herramientas de análisis de datos tradicionales.

b) Big Data no desafía la capacidad de las herramientas tradicionales de procesamiento de datos para capturar, almacenar, gestionar y analizar efectivamente la información, dada la evolución constante de las capacidades computacionales y algoritmos de optimización.

c) Aunque Big Data puede incluir datos estructurados, su naturaleza se expande al incorporar grandes cantidades de datos no estructurados y semiestructurados, lo que exige el uso de tecnologías especializadas en almacenamiento y procesamiento como Hadoop y sistemas de bases de datos NoSQL.

d) Además de su complejidad y diversidad, Big Data introduce desafíos significativos en términos de veracidad y variabilidad al requerir métodos avanzados de limpieza y validación de datos para asegurar la integridad del análisis.

Trabajo de Control
Sistemas de Recuperación de Información
Fecha: 15 de julio de 2024
Temario #61

Nombre: _____ Grupo: _____

Seleccione en cada caso la(s) respuesta(s) correcta(s). Considere también no marcar ninguna opción.

1. Implementar índices invertidos en un SRI asegura:
   a) Acelerar el proceso de búsqueda al permitir búsquedas directas por contenido en lugar de por título.
   b) Reducir la cantidad de espacio de almacenamiento necesario al comprimir los datos de los documentos.
   c) Facilitar la búsqueda sobre los datos que contienen términos específicos al mantener una lista de datos para cada término único.
   d) Incrementar la seguridad de los datos almacenados al dificultar el acceso directo a la información sin el índice correcto.

2. Dentro del ecosistema de Hadoop, el HDFS se caracteriza por:
   a) El modelo de acceso y de escritura de datos en tiempo real.
   b) La tolerancia a fallos mediante la replicación de datos.
   c) El almacenamiento exclusivo para archivos de texto.
   d) La capacidad ilimitada de almacenamiento.

3. Sobre WordNet y su aplicación en el procesamiento del lenguaje natural, se puede afirmar que:
   a) El diseño de WordNet facilita su integración en aplicaciones multilingües de NLP, aunque su desarrollo original se centró en el inglés.
   b) Aunque WordNet es una herramienta valiosa en el NLP, su estructura no incluye información sobre la frecuencia de uso de las palabras en el lenguaje natural.
   c) Los synsets facilitan la identificación de relaciones semánticas entre palabras, como la hiperonimia y la meronimia, enriqueciendo tareas de NLP.
   d) WordNet proporciona una base para la desambiguación semántica al agrupar palabras con significados similares en synsets.

4. Sobre el algoritmo de PageRank, visto en clase, se puede afirmar que:
   a) Evalúa la importancia de un sitio web en función de la calidad y cantidad de enlaces entrantes que recibe de otros sitios web.
   b) Solo tiene en cuenta el contenido en un sitio web para determinar su relevancia en los resultados de búsqueda.
   c) Asigna una puntuación alta a los sitios web que tienen un gran número de enlaces entrantes sin tener en cuenta la calidad de esos enlaces.
   d) Asigna una puntuación baja a los sitios web que contienen muchos enlaces salientes, ya que indica una falta de relevancia.

5. La computación evolutiva:
   a) No es aplicable en la RI debido a la complejidad de los algoritmos evolutivos.
   b) Solo puede manejar conjuntos de datos pequeños y no es escalable a grandes volúmenes de datos.
   c) Utiliza algoritmos para buscar soluciones óptimas en grandes espacios de búsqueda, lo que la hace adecuada para problemas de optimización en la RI.
   d) Es útil solo para problemas de clasificación de documentos y no para otras tareas de RI en general.

6. En la RI en el contexto de Big Data, se puede asegurar que:
   a) La indexación distribuida divide los datos en múltiples fragmentos que se almacenan en varios nodos para permitir búsquedas paralelas y mejorar la escalabilidad.
   b) La indexación distribuida no ofrece ventajas en términos de rendimiento y escalabilidad en comparación con la indexación centralizada.
   c) La indexación distribuida es una técnica obsoleta.
   d) La indexación distribuida almacena todos los datos en un solo servidor para facilitar su acceso y búsqueda.

7. La transición de la Web 1.0 a la Web 2.0 se caracterizó principalmente por:
   a) El aumento en la velocidad de conexión a internet, que permitió una mejor calidad de las páginas web.
   b) La disminución de la importancia de los motores de búsqueda en la navegación web.
   c) El cambio de páginas web estáticas a dinámicas, permitiendo la interacción del usuario y la generación de contenido.
   d) La reducción en el uso de HTML y CSS en el desarrollo de sitios web.

8. En el diseño de un videojuego de roles (RPG) en el que los personajes tienen atributos como salud, fuerza y velocidad, ¿cuál de las siguientes opciones representa mejor una implementación de la representación del conocimiento orientado a objetos?

   a) Cada personaje se representa como una lista de cadenas de texto que describen sus características físicas y habilidades.

   b) Cada personaje se representa como una función que calcula sus atributos en función de su nivel y experiencia.

   c) Cada personaje se representa como una matriz de números que almacena sus valores de atributos.

   d) Cada personaje se representa como un objeto con propiedades como salud, fuerza y velocidad, y métodos para modificar y consultar estos valores.

9. Si una red cumple la propiedad de ser un grafo de mundo pequeño, entonces se conoce que:

   a) El número de componentes fuertemente conexas está relacionado con la cantidad de grafos K_n presentes.

   b) La longitud media del camino entre todo par de vértices es pequeña.

   c) El grafo es un anillo regular de grado 5.

   d) La red posee pocos vértices.

10. Se puede afirmar que:

    a) No es posible identificar subgrupos dentro de una red utilizando análisis de redes.

    b) La cantidad de conexiones de un nodo siempre indica su influencia en la red.

    c) El tamaño de una red es siempre indicativo de su efectividad en la transmisión de información.

    d) Todas las relaciones en una red tienen la misma importancia para el análisis.

11. Se puede afirmar que:

    a) El término "Big Data" se refiere exclusivamente al volumen de datos que una organización maneja, sin tener en cuenta la velocidad, la variedad y la veracidad de los datos.

    b) Los SRI pueden beneficiarse de MapReduce para mejorar la RI relevante.

    c) MapReduce es un modelo de procesamiento distribuido utilizado para trabajar con grandes volúmenes de datos.

    d) Uno de los desafíos en el procesamiento de Big Data es la capacidad de gestionar y analizar datos provenientes de diversas fuentes y en diferentes formatos de manera eficiente.

12. En una red de transporte donde cada nodo es una parada de autobús y las aristas representan si existe un carro que pasa por ambos sitios, ¿qué puede mejorar el sistema de transporte?

    a) No considerar la centralidad de intermediación de las estaciones de transporte público al planificar rutas y horarios, ya que no tiene impacto en las conexiones entre las paradas.

    b) Utilizar el grafo inducido de los nodos con mayor valor en la centralidad de grado para aplicar la centralidad de intermediación con el propósito de reforzar las paradas con mayor tráfico.

    c) Utilizar la centralidad de intermediación para identificar las paradas de transferencia clave y establecer nuevas conexiones entre diferentes líneas de transporte público.

    d) No implementar sistemas de información en tiempo real para los usuarios, ya que pueden aumentar la carga de los trabajadores.

13. Una plataforma de comercio electrónico desea mejorar la experiencia del usuario al permitir una navegación más personalizada y contextualizada. Actualmente los usuarios tienen dificultades para encontrar productos relevantes debido a la gran cantidad de opciones disponibles. La empresa está interesada en implementar características de la Web 2.5 y la Web Semántica para abordar este problema. ¿Qué características podrían ayudar para ofrecer una navegación más personalizada y contextualizada?

    a) La implementación de ontologías y metadatos para enriquecer la descripción de productos y mejorar la precisión de las recomendaciones.

    b) La optimización de la velocidad de carga del sitio web para mejorar la experiencia del usuario y reducir el abandono del carrito de compra.

    c) La integración de redes sociales para permitir la recomendación de productos basada en las preferencias de amigos y contactos.

    d) La incorporación de los productos en tendencia en el mercado.

14. La diferencia entre la indexación por tokens y la indexación por conceptos puede definirse como:

    a) La indexación por tokens asigna pesos a los términos basados en su importancia relativa, mientras que la indexación por conceptos utiliza un sistema de etiquetado para asociar términos con características generales.

    b) La indexación por tokens divide los datos en términos individuales, mientras que la indexación por conceptos agrupa los datos en categorías definidas.

c) La indexación por tokens asigna un valor numérico a cada término de los datos, mientras que la indexación por conceptos utiliza algoritmos de encriptación para proteger la privacidad de los datos.

d) La indexación por tokens normaliza los datos reduciéndolos a su forma básica, mientras que la indexación por conceptos utiliza un método de ordenación para organizar los términos característicos de los datos.

15. Existen varias estrategias consideradas como no recomendables para mejorar el SEO de un sitio web. Dentro de estas estrategias negativas, se encuentran:

a) Construir enlaces naturales de sitios web con autoridad y relevancia temática.

b) Evitar el uso de las meta etiquetas y descripciones del sitio web para reflejar el contenido de la página.

c) Incluir una cantidad excesiva de palabras clave irrelevantes para intentar manipular los rankings de búsqueda.

d) Utilizar técnicas de encubrimiento para mostrar contenido diferente a lo indexado por los motores de búsqueda.

16. El uso de Hadoop y MapReduce en el contexto de la RI tiene como objetivo:

a) Automatizar el mantenimiento de sistemas de bases de datos.

b) Facilitar el análisis en tiempo real de datos de redes sociales.

c) Mejorar la eficiencia energética en centros de datos.

d) Procesar y analizar grandes conjuntos de datos para la RI.

17. El concepto de *rank sink* en el algoritmo de PageRank representa páginas web:

a) Con distribuciones uniformes de la puntuación de PageRank.

b) Que tienen una puntuación de PageRank más alta que otras debido a la manipulación de enlaces entrantes y salientes.

c) Con una baja calidad de contenido y una cantidad insuficiente de enlaces salientes, lo que las hace menos relevantes en los resultados de búsqueda.

d) Con un alto número de enlaces salientes que no reciben enlaces entrantes, lo que puede afectar negativamente su puntuación de PageRank.

18. Al "relajar" el concepto de clique en la detección de comunidades se intenta solucionar:

a) La imposibilidad de conectar nodos distantes.

b) La necesidad de datos externos para analizar la red.

c) La uniformidad de los nodos en términos de grado.

d) El solapamiento y la complejidad computacional.

19. Una buena práctica de SEO para mejorar el posicionamiento de un sitio web en los motores de búsqueda es:

a) Obtener enlaces de otros sitios web relevantes y de calidad que apunten al sitio.

b) Copiar contenido directamente de otros sitios web populares para aumentar la cantidad de páginas indexadas.

c) Utilizar técnicas de encubrimiento para mostrar contenido diferente a lo indexado por los motores de búsqueda.

d) Llenar el contenido con palabras clave irrelevantes para aumentar la densidad de palabras clave.

20. Un investigador necesita recopilar datos de múltiples sitios web para un estudio académico, pero se enfrenta a varios desafíos al realizar el proceso de Web Scraping de manera ética y legal. ¿Cuál de las siguientes opciones describe mejor uno de los desafíos asociados al proceso de Web Scraping?

a) La necesidad de comprender la estructura del sitio web y su código HTML para extraer los datos correctamente.

b) La disponibilidad limitada de datos en línea que se pueden extraer utilizando técnicas de Web Scraping.

c) La dificultad para encontrar herramientas de Web Scraping gratuitas y fiables.

d) La necesidad de estar montado sobre un scrawler que cumpla con todas las políticas.

Trabajo de Control
Sistemas de Recuperación de Información
Fecha: 15 de julio de 2024
Temario #62

Nombre: _____ Grupo: _____

Seleccione en cada caso la(s) respuesta(s) correcta(s). Considere también no marcar ninguna opción.

1. Implementar índices invertidos en un SRI asegura:
   a) Acelerar el proceso de búsqueda al permitir búsquedas directas por contenido en lugar de por título.
   b) Reducir la cantidad de espacio de almacenamiento necesario al comprimir los datos de los documentos.
   c) Facilitar la búsqueda sobre los datos que contienen términos específicos al mantener una lista de datos para cada término único.
   d) Incrementar la seguridad de los datos almacenados al dificultar el acceso directo a la información sin el índice correcto.

2. Si se tiene el conjunto de páginas interconectadas $\{A \to B, C; B \to C; C \to A\}$, entonces la página con valor más alto de PageRank es:
   a) A y B.
   b) A.
   c) C.
   d) B.

3. ¿Cuál de las siguientes opciones describe mejor la diferencia clave entre Web Crawling y Web Scraping?
   a) Web Crawling se centra en la exploración y recopilación de enlaces de múltiples sitios web, mientras que Web Scraping se enfoca en la extracción específica de datos de páginas web individuales.
   b) Web Crawling se realiza utilizando herramientas de automatización como Selenium WebDriver, mientras que Web Scraping se lleva a cabo mediante el análisis de HTML y CSS.
   c) Web Scraping implica el análisis de la estructura y el contenido de las páginas web para extraer datos, mientras que Web Crawling se refiere a la descarga y almacenamiento de páginas web completas.
   d) Web Scraping es más eficaz para rastrear e indexar contenido web para motores de búsqueda, mientras que Web Crawling se utiliza principalmente para la extracción de datos en proyectos de investigación.

4. Se puede afirmar que:
   a) El término "Big Data" se refiere exclusivamente al volumen de datos que una organización maneja, sin tener en cuenta la velocidad, la variedad y la veracidad de los datos.
   b) Los SRI pueden beneficiarse de MapReduce para mejorar la RI relevante.
   c) MapReduce es un modelo de procesamiento distribuido utilizado para trabajar con grandes volúmenes de datos.
   d) Uno de los desafíos en el procesamiento de Big Data es la capacidad de gestionar y analizar datos provenientes de diversas fuentes y en diferentes formatos de manera eficiente.

5. Se tiene un grafo donde cada nodo es un personaje de cierto libro de cuentos y la existencia de las aristas está definida si dos personajes aparecen en el mismo cuento. Se puede asegurar que:
   a) La centralidad de intermediación identifica a los personajes que actúan como conectores entre personajes de cuentos distintos.
   b) La centralidad de grado es útil para identificar los personajes con más conexiones dentro de la red, lo que puede indicar su importancia en el libro.
   c) La centralidad de cercanía ofrece una relación entre la cantidad de vecinos de un nodo con respecto a la longitud máxima de un camino dentro del grafo partiendo del nodo en cuestión.
   d) La centralidad de cercanía indica el grado de conexión de cada personaje con el resto de los personajes de los cuentos del libro.

6. ¿Qué algoritmos permiten obtener información de una red?
   a) Índices de centralidad.
   b) Detección de comunidades.
   c) Hypertext Induced Topic Selection (HITS).
   d) PageRank.

7. Analizar una red permite:
   a) Detectar posibles tendencias antes de que se conviertan en tendencia.
   b) Obtener predicciones exactas de eventos futuros en mercados financieros.
   c) Evaluar la calidad del contenido de un sitio web.
   d) Encontrar nodos "sensibles" o críticos para la red.

8. En la RI en el contexto de Big Data, se puede asegurar que:

a) La indexación distribuida divide los datos en múltiples fragmentos que se almacenan en varios nodos para permitir búsquedas paralelas y mejorar la escalabilidad.

b) La indexación distribuida no ofrece ventajas en términos de rendimiento y escalabilidad en comparación con la indexación centralizada.

c) La indexación distribuida es una técnica obsoleta.

d) La indexación distribuida almacena todos los datos en un solo servidor para facilitar su acceso y búsqueda.

9. La centralidad de intermediación de un nodo indica:

a) La cantidad de nodos vecinos directos.

b) La resistencia del nodo a fallos.

c) El número total de conexiones entrantes y salientes.

d) La frecuencia con la que un nodo actúa como puente en el camino más corto entre otros dos nodos.

10. En un sistema de control de tráfico urbano basado en reglas, ¿cuál de las siguientes reglas sería más efectiva para manejar situaciones de congestión en una intersección?

a) Si hay pocos vehículos en la intersección, reducir el tiempo de los semáforos en verde.

b) Si hay muchos vehículos en la intersección, aumentar el tiempo de los semáforos en verde.

c) Si hay un vehículo de emergencia en la intersección, detener todos los demás vehículos.

d) Si hay muchos peatones cruzando la intersección, reducir el tiempo de los semáforos en rojo.

11. Para transformar el contenido no estructurado de las páginas web en datos estructurados, el Web Scraping:

a) Extrae información basada en patrones de HTML/CSS.

b) Analiza los protocolos de red.

c) Interpreta el código JavaScript en tiempo real.

d) Convierte de forma automática imágenes a texto.

12. Se puede afirmar que:

a) No es posible identificar subgrupos dentro de una red utilizando análisis de redes.

b) La cantidad de conexiones de un nodo siempre indica su influencia en la red.

c) El tamaño de una red es siempre indicativo de su efectividad en la transmisión de información.

d) Todas las relaciones en una red tienen la misma importancia para el análisis.

13. La computación evolutiva:

a) No es aplicable en la RI debido a la complejidad de los algoritmos evolutivos.

b) Solo puede manejar conjuntos de datos pequeños y no es escalable a grandes volúmenes de datos.

c) Utiliza algoritmos para buscar soluciones óptimas en grandes espacios de búsqueda, lo que la hace adecuada para problemas de optimización en la RI.

d) Es útil solo para problemas de clasificación de documentos y no para otras tareas de RI en general.

14. La premisa básica del algoritmo de PageRank para clasificar páginas web en los resultados de búsqueda es:

a) La antigüedad de la página web es el principal factor para determinar su clasificación.

b) El contenido y la relevancia de las palabras clave en la página web determinan su posición.

c) Los enlaces entrantes a una página web desde otras páginas contribuyen a su importancia y clasificación.

d) La cantidad de visitas que recibe una página web determina su clasificación.

15. Considerando las prácticas éticas y legales en el Web Scraping, se puede asegurar que:

a) Web Scraping requiere considerar las políticas de `robots.txt` del sitio web objetivo.

b) Es importante revisar y respetar los términos de servicio del sitio web, así como las leyes aplicables de protección de datos y derechos de autor, antes de realizar Web Scraping.

c) Web Scraping sobre datos personales sin consentimiento es generalmente aceptado si los datos se utilizan con fines de investigación.

d) La extracción de datos mediante Web Scraping siempre es legal, independientemente de las leyes locales sobre derechos de autor y privacidad de datos definidos en los sitios web.

16. El concepto de *rank sink* en el algoritmo de PageRank representa páginas web:

a) Con distribuciones uniformes de la puntuación de PageRank.

b) Que tienen una puntuación de PageRank más alta que otras debido a la manipulación de enlaces entrantes y salientes.

c) Con una baja calidad de contenido y una cantidad insuficiente de enlaces salientes, lo que las hace menos relevantes en los resultados de búsqueda.

d) Con un alto número de enlaces salientes que no reciben enlaces entrantes, lo que puede afectar negativamente su puntuación de PageRank.

17. En el contexto del procesamiento de grandes conjuntos de datos, se puede asegurar que:

a) MapReduce es un enfoque para procesar datos de forma secuencial en un solo servidor para evitar problemas de concurrencia.

b) MapReduce divide una tarea en múltiples pasos de map y reduce que se ejecutan de forma secuencial en diferentes servidores para mejorar el rendimiento y la escalabilidad.

c) MapReduce no es adecuado para el procesamiento de datos no estructurados.

d) MapReduce solo puede manejar pequeñas cantidades de datos y no escala bien a grandes conjuntos de datos.

18. La afirmación que mejor refleja el principio subyacente de PageRank, considerando su importancia en la clasificación de los sitios web, es:

a) PageRank valora más la cantidad de enlaces entrantes a una página web, independientemente de la calidad o relevancia de estos enlaces.

b) La efectividad de PageRank se basa exclusivamente en el análisis de las palabras clave contenidas en los enlaces entrantes, sin considerar la estructura de enlace de la Web.

c) El algoritmo de PageRank considera tanto la cantidad como la calidad de los enlaces entrantes, asignando mayor valor a los enlaces provenientes de sitios web considerados como "importantes".

d) PageRank opera bajo el supuesto de que los enlaces entrantes y salientes tienen el mismo impacto en la valoración de la relevancia de una página web.

19. Al "relajar" el concepto de clique en la detección de comunidades se intenta solucionar:

a) La imposibilidad de conectar nodos distantes.

b) La necesidad de datos externos para analizar la red.

c) La uniformidad de los nodos en términos de grado.

d) El solapamiento y la complejidad computacional.

20. La diferencia entre la indexación por tokens y la indexación por conceptos puede definirse como:

a) La indexación por tokens asigna pesos a los términos basados en su importancia relativa, mientras que la indexación por conceptos utiliza un sistema de etiquetado para asociar términos con características generales.

b) La indexación por tokens divide los datos en términos individuales, mientras que la indexación por conceptos agrupa los datos en categorías definidas.

c) La indexación por tokens asigna un valor numérico a cada término de los datos, mientras que la indexación por conceptos utiliza algoritmos de encriptación para proteger la privacidad de los datos.

d) La indexación por tokens normaliza los datos reduciéndolos a su forma básica, mientras que la indexación por conceptos utiliza un método de ordenación para organizar los términos característicos de los datos.

Trabajo de Control
Sistemas de Recuperación de Información
Fecha: 15 de julio de 2024
Temario #63

Nombre: _____ Grupo: _____

Seleccione en cada caso la(s) respuesta(s) correcta(s). Considere también no marcar ninguna opción.

1. El concepto de *rank sink* en el algoritmo de PageRank representa páginas web:
   a) Con distribuciones uniformes de la puntuación de PageRank.
   b) Que tienen una puntuación de PageRank más alta que otras debido a la manipulación de enlaces entrantes y salientes.
   c) Con una baja calidad de contenido y una cantidad insuficiente de enlaces salientes, lo que las hace menos relevantes en los resultados de búsqueda.
   d) Con un alto número de enlaces salientes que no reciben enlaces entrantes, lo que puede afectar negativamente su puntuación de PageRank.

2. El conjunto frontera de URLs en un Web Crawler:
   a) Delimita los sitios web que visitará en futuras iteraciones del proceso.
   b) Es similar a un conjunto de URLs que esperan ser visitadas.
   c) Almacena hipervínculos que pertenecen al mismo dominio del conjunto semilla de URLs con que inició el crawler.
   d) Indica las secciones que no pueden ser visitadas de cada sitio web.

3. El uso de Hadoop y MapReduce en el contexto de la RI tiene como objetivo:
   a) Automatizar el mantenimiento de sistemas de bases de datos.
   b) Facilitar el análisis en tiempo real de datos de redes sociales.
   c) Mejorar la eficiencia energética en centros de datos.
   d) Procesar y analizar grandes conjuntos de datos para la RI.

4. Los algoritmos para detectar comunidades en una red intentan:
   a) Buscar conjuntos donde cada nodo de un mismo conjunto tenga características similares al resto de los nodos del conjunto.
   b) Seleccionar aleatoriamente nodos de alto grado y sus vecinos.
   c) Buscar subgrafos tal que no incluyan nodos cuya ausencia desconecte al subgrafo.
   d) Encontrar grupos donde los nodos pertenecientes a los mismos grupos son cercanos bajo cierta métrica y lejanos con respecto a los nodos de otros grupos.

5. En un sistema donde el conocimiento está definido a partir de reglas se puede asegurar que:
   a) El nivel de especificidad de las reglas está limitado.
   b) La representación de conocimiento está basada en la lógica proposicional.
   c) El razonamiento se activa en cadena hacia delante.
   d) El orden en que se definen las reglas no altera el razonamiento del sistema.

6. Referente al Web Crawler puede afirmarse que:
   a) Los hipervínculos encontrados en cada sitio web que no pertenecen al dominio donde fueron encontrados se desechan, puesto que no expande el conjunto de URLs sin visitar.
   b) Las páginas visitadas no se procesan nunca más.
   c) No necesita de un conjunto inicial de URLs para recorrer la Web.
   d) No tiene como objetivo indexar y recopilar información de diferentes sitios web.

7. La centralidad de intermediación de un nodo indica:
   a) La cantidad de nodos vecinos directos.
   b) La resistencia del nodo a fallos.
   c) El número total de conexiones entrantes y salientes.
   d) La frecuencia con la que un nodo actúa como puente en el camino más corto entre otros dos nodos.

8. La integración de Hadoop y MapReduce en la RI trae como ventaja:
   a) La eliminación de la necesidad de sistemas de bases de datos.
   b) La posibilidad de la extracción de información relevante y la generación de resultados significativos de grandes conjuntos de datos.
   c) La garantía de la privacidad absoluta de los datos procesados.
   d) La reducción de los costos operativos a cero.

9. La Web 3 se conoce como:
   a) Internet de las cosas.
   b) Web de solo lectura.

c) Web semántica.

d) Web de escritura-lectura.

10. Existen varias estrategias consideradas como no recomendables para mejorar el SEO de un sitio web. Dentro de estas estrategias negativas, se encuentran:
    a) Construir enlaces naturales de sitios web con autoridad y relevancia temática.
    b) Evitar el uso de las meta etiquetas y descripciones del sitio web para reflejar el contenido de la página.
    c) Incluir una cantidad excesiva de palabras clave irrelevantes para intentar manipular los rankings de búsqueda.
    d) Utilizar técnicas de encubrimiento para mostrar contenido diferente a lo indexado por los motores de búsqueda.

11. No se considera como técnica para detectar comunidades en una red:
    a) Analizar la mutualidad de los enlaces.
    b) Usar el agrupamiento jerárquico.
    c) Utilizar el algoritmo de K-Means.
    d) Encontrar cliques de vértices de grado par.

12. En el diseño de un videojuego de roles (RPG) en el que los personajes tienen atributos como salud, fuerza y velocidad, ¿cuál de las siguientes opciones representa mejor una implementación de la representación del conocimiento orientado a objetos?
    a) Cada personaje se representa como una lista de cadenas de texto que describen sus características físicas y habilidades.
    b) Cada personaje se representa como una función que calcula sus atributos en función de su nivel y experiencia.
    c) Cada personaje se representa como una matriz de números que almacena sus valores de atributos.
    d) Cada personaje se representa como un objeto con propiedades como salud, fuerza y velocidad, y métodos para modificar y consultar estos valores.

13. La web actual se enfrenta a problemas como:
    a) Presencia de grandes volúmenes de datos estructurados.
    b) Presencia de una alta calidad en los datos.
    c) Presencia de datos volátiles y distribuidos.
    d) Heterogeneidad en los datos.

14. ¿Qué algoritmos permiten obtener información de una red?
    a) Índices de centralidad.
    b) Detección de comunidades.
    c) Hypertext Induced Topic Selection (HITS).
    d) PageRank.

15. Si una red cumple la propiedad de ser un grafo de mundo pequeño, entonces se conoce que:
    a) El número de componentes fuertemente conexas está relacionado con la cantidad de grafos K_n presentes.
    b) La longitud media del camino entre todo par de vértices es pequeña.
    c) El grafo es un anillo regular de grado 5.
    d) La red posee pocos vértices.

16. La premisa básica del algoritmo de PageRank para clasificar páginas web en los resultados de búsqueda es:
    a) La antigüedad de la página web es el principal factor para determinar su clasificación.
    b) El contenido y la relevancia de las palabras clave en la página web determinan su posición.
    c) Los enlaces entrantes a una página web desde otras páginas contribuyen a su importancia y clasificación.
    d) La cantidad de visitas que recibe una página web determina su clasificación.

17. La política de ordenación de URLs en los Web Crawlers tiene como aspecto fundamental:
    a) Limitar el acceso de los crawlers a ciertas secciones de un sitio web, evitando el rastreo de URLs consideradas menos importantes o sensibles.
    b) Definir la estructura de la URL de destino, asegurando que estén ordenadas alfabéticamente para facilitar la navegación y la indexación.
    c) Establecer la prioridad de rastreo de las URLs, determinando el orden en que los crawlers visitan y procesan cada página web.
    d) Determinar la forma en que los crawlers asignan un valor de relevancia a cada URL para clasificarlas en el índice de búsqueda.

18. Los Web Crawlers se enfrentan a desafíos constantes. Dentro de ellos se encuentran:
    a) La dificultad para generar el contenido dinámico en tiempo real.
    b) La modificación del código y la estructura del sitio web.
    c) La incapacidad para interpretar correctamente el lenguaje de programación utilizado en el desarrollo de los sitios web.
    d) La falta de acceso a la base de datos del servidor web para extraer información actualizada.

19. El componente responsable de la gestión de recursos y planificación de tareas en Hadoop es:
    a) MapReduce.

b) Hadoop Common.

c) HDFS.

d) YARN.

20. En el contexto de la RI en redes, se puede afirmar que:

a) La detección de comunidades no es relevante para la RI en redes, ya que se centra únicamente en la estructura de la red sin considerar el contenido.

b) La detección de comunidades en una red siempre produce resultados objetivos y consistentes independientemente del algoritmo utilizado.

c) La detección de comunidades ayuda a identificar grupos de nodos altamente conectados entre sí, lo que puede ser útil para comprender la estructura y el contenido de la red.

d) La detección de comunidades solo se aplica a redes pequeñas y simples, no a redes grandes y complejas.

Trabajo de Control
Sistemas de Recuperación de Información
Fecha: 15 de julio de 2024
Temario #64

Nombre: _____ Grupo: _____

Seleccione en cada caso la(s) respuesta(s) correcta(s). Considere también no marcar ninguna opción.

1. La Web 1.0 se caracteriza por:
   a) Las actualizaciones del contenido de los sitios web se efectúa de forma manual dentro del código HTML
   b) Los sitios web se centran en brindar información en lugar de facilitar la colaboración o participación de los usuarios.
   c) Uso de comunidades virtuales para popularizar los sitios web de noticias.
   d) Los propietarios de los sitios web proporcionan contenido de forma periódica.

2. En la RI en el contexto de Big Data, se puede asegurar que:
   a) La indexación distribuida divide los datos en múltiples fragmentos que se almacenan en varios nodos para permitir búsquedas paralelas y mejorar la escalabilidad.
   b) La indexación distribuida no ofrece ventajas en términos de rendimiento y escalabilidad en comparación con la indexación centralizada.
   c) La indexación distribuida es una técnica obsoleta.
   d) La indexación distribuida almacena todos los datos en un solo servidor para facilitar su acceso y búsqueda.

3. Una buena práctica de SEO para mejorar el posicionamiento de un sitio web en los motores de búsqueda es:
   a) Obtener enlaces de otros sitios web relevantes y de calidad que apunten al sitio.
   b) Copiar contenido directamente de otros sitios web populares para aumentar la cantidad de páginas indexadas.
   c) Utilizar técnicas de encubrimiento para mostrar contenido diferente a lo indexado por los motores de búsqueda.
   d) Llenar el contenido con palabras clave irrelevantes para aumentar la densidad de palabras clave.

4. Para transformar el contenido no estructurado de las páginas web en datos estructurados, el Web Scraping:
   a) Extrae información basada en patrones de HTML/CSS.
   b) Analiza los protocolos de red.
   c) Interpreta el código JavaScript en tiempo real.
   d) Convierte de forma automática imágenes a texto.

5. Se tiene un grafo donde cada nodo es un personaje de cierto libro de cuentos y la existencia de las aristas está definida si dos personajes aparecen en el mismo cuento. Se puede asegurar que:
   a) La centralidad de intermediación identifica a los personajes que actúan como conectores entre personajes de cuentos distintos.
   b) La centralidad de grado es útil para identificar los personajes con más conexiones dentro de la red, lo que puede indicar su importancia en el libro.
   c) La centralidad de cercanía ofrece una relación entre la cantidad de vecinos de un nodo con respecto a la longitud máxima de un camino dentro del grafo partiendo del nodo en cuestión.
   d) La centralidad de cercanía indica el grado de conexión de cada personaje con el resto de los personajes de los cuentos del libro.

6. Un investigador necesita recopilar datos de múltiples sitios web para un estudio académico, pero se enfrenta a varios desafíos al realizar el proceso de Web Scraping de manera ética y legal. ¿Cuál de las siguientes opciones describe mejor uno de los desafíos asociados al proceso de Web Scraping?
   a) La necesidad de comprender la estructura del sitio web y su código HTML para extraer los datos correctamente.
   b) La disponibilidad limitada de datos en línea que se pueden extraer utilizando técnicas de Web Scraping.
   c) La dificultad para encontrar herramientas de Web Scraping gratuitas y fiables.
   d) La necesidad de estar montado sobre un scrawler que cumpla con todas las políticas.

7. Un sistema cuenta con la siguiente información:
   - Juan pertenece a la marina.
   - Juan es capellán (sacerdote encargado de una tarea específica fuera de la parroquia).
   - Los infantes de la marina suelen ser bebedores de cerveza.
   - Un capellán no suele ser bebedor de cerveza.
   - Un bebedor de cerveza suele tener sobrepeso.

- Por lo general, un infante de la marina está en buena condición física.

Si se representa la información en una red de herencia se puede concluir que:

a) No se puede asegurar que Juan sea capellán.

b) No se puede asegurar que Juan esté en buena condición física.

c) Juan tiene sobrepeso.

d) Existen dos razonamientos cancelables.

8. En el modelo de representación del conocimiento basado en herencia se puede asegurar que:

a) La cantidad de padres que puede tener un nodo no es mayor que 4.

b) Las conclusiones pueden ser canceladas si el grafo es ambiguo.

c) Solo se usa en entornos referentes a la biología.

d) El razonamiento deducido está respaldado por al menos un camino dentro del grafo.

9. La diferencia entre la indexación por tokens y la indexación por conceptos puede definirse como:

a) La indexación por tokens asigna pesos a los términos basados en su importancia relativa, mientras que la indexación por conceptos utiliza un sistema de etiquetado para asociar términos con características generales.

b) La indexación por tokens divide los datos en términos individuales, mientras que la indexación por conceptos agrupa los datos en categorías definidas.

c) La indexación por tokens asigna un valor numérico a cada término de los datos, mientras que la indexación por conceptos utiliza algoritmos de encriptación para proteger la privacidad de los datos.

d) La indexación por tokens normaliza los datos reduciéndolos a su forma básica, mientras que la indexación por conceptos utiliza un método de ordenación para organizar los términos característicos de los datos.

10. La integración de Hadoop y MapReduce en la RI trae como ventaja:

a) La eliminación de la necesidad de sistemas de bases de datos.

b) La posibilidad de la extracción de información relevante y la generación de resultados significativos de grandes conjuntos de datos.

c) La garantía de la privacidad absoluta de los datos procesados.

d) La reducción de los costos operativos a cero.

11. En una red de transporte donde cada nodo es una parada de autobús y las aristas representan si existe un carro que pasa por ambos sitios, ¿qué puede mejorar el sistema de transporte?

a) No considerar la centralidad de intermediación de las estaciones de transporte público al planificar rutas y horarios, ya que no tiene impacto en las conexiones entre las paradas.

b) Utilizar el grafo inducido de los nodos con mayor valor en la centralidad de grado para aplicar la centralidad de intermediación con el propósito de reforzar las paradas con mayor tráfico.

c) Utilizar la centralidad de intermediación para identificar las paradas de transferencia clave y establecer nuevas conexiones entre diferentes líneas de transporte público.

d) No implementar sistemas de información en tiempo real para los usuarios, ya que pueden aumentar la carga de los trabajadores.

12. El conjunto frontera de URLs en un Web Crawler:

a) Delimita los sitios web que visitará en futuras iteraciones del proceso.

b) Es similar a un conjunto de URLs que esperan ser visitadas.

c) Almacena hipervínculos que pertenecen al mismo dominio del conjunto semilla de URLs con que inició el crawler.

d) Indica las secciones que no pueden ser visitadas de cada sitio web.

13. En el diseño de un videojuego de roles (RPG) en el que los personajes tienen atributos como salud, fuerza y velocidad, ¿cuál de las siguientes opciones representa mejor una implementación de la representación del conocimiento orientado a objetos?

a) Cada personaje se representa como una lista de cadenas de texto que describen sus características físicas y habilidades.

b) Cada personaje se representa como una función que calcula sus atributos en función de su nivel y experiencia.

c) Cada personaje se representa como una matriz de números que almacena sus valores de atributos.

d) Cada personaje se representa como un objeto con propiedades como salud, fuerza y velocidad, y métodos para modificar y consultar estos valores.

14. No se considera como técnica para detectar comunidades en una red:

a) Analizar la mutualidad de los enlaces.

b) Usar el agrupamiento jerárquico.

c) Utilizar el algoritmo de K-Means.

d) Encontrar cliques de vértices de grado par.

15. Para la RI, el análisis de las redes puede:

a) Ayudar a identificar grupos de interés.

b) Indicar la importancia de una entidad en la transmisión de la información.

c) Ayudar a comprender la conectividad y la accesibilidad entre las entidades.

d) Revelar patrones de influencia dentro de una comunidad.

16. Sobre el algoritmo de PageRank, visto en clase, se puede afirmar que:

a) Evalúa la importancia de un sitio web en función de la calidad y cantidad de enlaces entrantes que recibe de otros sitios web.

b) Solo tiene en cuenta el contenido en un sitio web para determinar su relevancia en los resultados de búsqueda.

c) Asigna una puntuación alta a los sitios web que tienen un gran número de enlaces entrantes sin tener en cuenta la calidad de esos enlaces.

d) Asigna una puntuación baja a los sitios web que contienen muchos enlaces salientes, ya que indica una falta de relevancia.

17. La Web 3 se conoce como:

a) Internet de las cosas.

b) Web de solo lectura.

c) Web semántica.

d) Web de escritura-lectura.

18. En un sistema donde el conocimiento está orientado a objetos se puede asegurar que:

a) Se enfatiza la atención a la información de la cual se extrajo el conocimiento.

b) Los marcos y las bandas son las estructuras utilizadas para la representación del modelo.

c) No es posible definir especializaciones de los objetos de la vida real dentro del sistema.

d) Las funciones de agregación dificultan poder establecer relaciones entre los objetos.

19. En el contexto del procesamiento de grandes conjuntos de datos, se puede asegurar que:

a) MapReduce es un enfoque para procesar datos de forma secuencial en un solo servidor para evitar problemas de concurrencia.

b) MapReduce divide una tarea en múltiples pasos de map y reduce que se ejecutan de forma secuencial en diferentes servidores para mejorar el rendimiento y la escalabilidad.

c) MapReduce no es adecuado para el procesamiento de datos no estructurados.

d) MapReduce solo puede manejar pequeñas cantidades de datos y no escala bien a grandes conjuntos de datos.

20. El componente responsable de la gestión de recursos y planificación de tareas en Hadoop es:

a) MapReduce.

b) Hadoop Common.

c) HDFS.

d) YARN.

Trabajo de Control
Sistemas de Recuperación de Información
Fecha: 15 de julio de 2024
Temario #65

Nombre: _____ Grupo: _____

Seleccione en cada caso la(s) respuesta(s) correcta(s). Considere también no marcar ninguna opción.

1. Al "relajar" el concepto de clique en la detección de comunidades se intenta solucionar:
   a) La imposibilidad de conectar nodos distantes.
   b) La necesidad de datos externos para analizar la red.
   c) La uniformidad de los nodos en términos de grado.
   d) El solapamiento y la complejidad computacional.

2. ¿Qué es un "Uniform Resource Locator (URL) Frontier" en el contexto de Web Crawling?
   a) Un protocolo que define cómo se deben formatear las URLs para el crawling.
   b) Una lista prioritaria de URLs que aún no han sido visitadas por el crawler.
   c) Una técnica para filtrar URLs irrelevantes y mejorar la eficiencia del crawling.
   d) Una base de datos que almacena URLs únicas identificadas como recursos en la web.

3. Los Web Crawlers se enfrentan a desafíos constantes. Dentro de ellos se encuentran:
   a) La dificultad para generar el contenido dinámico en tiempo real.
   b) La modificación del código y la estructura del sitio web.
   c) La incapacidad para interpretar correctamente el lenguaje de programación utilizado en el desarrollo de los sitios web.
   d) La falta de acceso a la base de datos del servidor web para extraer información actualizada.

4. Una empresa de comercio electrónico necesita mejorar su motor de búsqueda para proporcionar resultados más relevantes a sus usuarios. Actualmente, los resultados de la búsqueda no son precisos y los usuarios a menudo encuentran dificultades para encontrar productos específicos. La empresa está considerando implementar una indexación por conceptos para mejorar la relevancia de los resultados de búsqueda. ¿Qué beneficios podría provocar este cambio?
   a) Permite adaptarse fácilmente a cambios en el vocabulario y la terminología utilizada en las descripciones de los productos.
   b) Ayuda a identificar automáticamente relaciones entre productos, lo que puede mejorar las recomendaciones personalizadas a los usuarios.
   c) Facilita la visualización de productos al agruparlos por categorías o características comunes.
   d) Mejora la precisión en la búsqueda de productos relacionados, incluso cuando no coinciden exactamente con los términos de búsqueda del usuario.

5. Dentro del análisis de redes, la centralidad de grado mide la importancia de un nodo basándose en:
   a) El grado del nodo.
   b) La cantidad de veces que aparece el nodo en el camino mínimo entre cualquier par de nodos.
   c) La cantidad de vecinos del nodo.
   d) El número de aristas que posee el nodo.

6. La integración de Hadoop y MapReduce en la RI trae como ventaja:
   a) La eliminación de la necesidad de sistemas de bases de datos.
   b) La posibilidad de la extracción de información relevante y la generación de resultados significativos de grandes conjuntos de datos.
   c) La garantía de la privacidad absoluta de los datos procesados.
   d) La reducción de los costos operativos a cero.

7. En la RI en el contexto de Big Data, se puede asegurar que:
   a) La indexación distribuida divide los datos en múltiples fragmentos que se almacenan en varios nodos para permitir búsquedas paralelas y mejorar la escalabilidad.
   b) La indexación distribuida no ofrece ventajas en términos de rendimiento y escalabilidad en comparación con la indexación centralizada.
   c) La indexación distribuida es una técnica obsoleta.
   d) La indexación distribuida almacena todos los datos en un solo servidor para facilitar su acceso y búsqueda.

8. En el contexto de la RI en redes sociales se puede afirmar que:
   a) El análisis de redes sociales se centra exclusivamente en la cantidad de seguidores que tiene un usuario en particular para determinar su influencia en la red.

b) El análisis de centralidad de intermediación se utiliza para identificar usuarios que son importantes en una red social debido a su posición como "puentes" entre diferentes grupos de usuarios.

c) El análisis de sentimientos se utiliza para determinar la popularidad de una publicación en redes sociales sin tener en cuenta la opinión de los usuarios.

d) El análisis de redes sociales no es útil para comprender la difusión de información en una red social específica, para ello se utiliza la medida de centralidad de vector propio.

9. ¿Qué algoritmos permiten obtener información de una red?

a) Índices de centralidad.

b) Detección de comunidades.

c) Hypertext Induced Topic Selection (HITS).

d) PageRank.

10. Un investigador necesita recopilar datos de múltiples sitios web para un estudio académico, pero se enfrenta a varios desafíos al realizar el proceso de Web Scraping de manera ética y legal. ¿Cuál de las siguientes opciones describe mejor uno de los desafíos asociados al proceso de Web Scraping?

a) La necesidad de comprender la estructura del sitio web y su código HTML para extraer los datos correctamente.

b) La disponibilidad limitada de datos en línea que se pueden extraer utilizando técnicas de Web Scraping.

c) La dificultad para encontrar herramientas de Web Scraping gratuitas y fiables.

d) La necesidad de estar montado sobre un scrawler que cumpla con todas las políticas.

11. Sobre WordNet y su aplicación en el procesamiento del lenguaje natural, se puede afirmar que:

a) El diseño de WordNet facilita su integración en aplicaciones multilingües de NLP, aunque su desarrollo original se centró en el inglés.

b) Aunque WordNet es una herramienta valiosa en el NLP, su estructura no incluye información sobre la frecuencia de uso de las palabras en el lenguaje natural.

c) Los synsets facilitan la identificación de relaciones semánticas entre palabras, como la hiperonimia y la meronimia, enriqueciendo tareas de NLP.

d) WordNet proporciona una base para la desambiguación semántica al agrupar palabras con significados similares en synsets.

12. Implementar índices invertidos en un SRI asegura:

a) Acelerar el proceso de búsqueda al permitir búsquedas directas por contenido en lugar de por título.

b) Reducir la cantidad de espacio de almacenamiento necesario al comprimir los datos de los documentos.

c) Facilitar la búsqueda sobre los datos que contienen términos específicos al mantener una lista de datos para cada término único.

d) Incrementar la seguridad de los datos almacenados al dificultar el acceso directo a la información sin el índice correcto.

13. El conjunto frontera de URLs en un Web Crawler:

a) Delimita los sitios web que visitará en futuras iteraciones del proceso.

b) Es similar a un conjunto de URLs que esperan ser visitadas.

c) Almacena hipervínculos que pertenecen al mismo dominio del conjunto semilla de URLs con que inició el crawler.

d) Indica las secciones que no pueden ser visitadas de cada sitio web.

14. Para la RI, el análisis de las redes puede:

a) Ayudar a identificar grupos de interés.

b) Indicar la importancia de una entidad en la transmisión de la información.

c) Ayudar a comprender la conectividad y la accesibilidad entre las entidades.

d) Revelar patrones de influencia dentro de una comunidad.

15. El posicionamiento de un sitio web en los motores de búsqueda puede ser afectado por:

a) Utilizar técnicas de encubrimiento para mostrar contenido diferente a lo indexado por los motores de búsqueda.

b) Contenido sin valor en el sitio web.

c) Mantener una estructura de URL clara y coherente.

d) Obtener enlaces de sitios web irrelevantes y de baja calidad.

16. Una buena práctica de SEO para mejorar el posicionamiento de un sitio web en los motores de búsqueda es:

a) Obtener enlaces de otros sitios web relevantes y de calidad que apunten al sitio.

b) Copiar contenido directamente de otros sitios web populares para aumentar la cantidad de páginas indexadas.

c) Utilizar técnicas de encubrimiento para mostrar contenido diferente a lo indexado por los motores de búsqueda.

d) Llenar el contenido con palabras clave irrelevantes para aumentar la densidad de palabras clave.

17. La computación evolutiva:

    a) No es aplicable en la RI debido a la complejidad de los algoritmos evolutivos.

    b) Solo puede manejar conjuntos de datos pequeños y no es escalable a grandes volúmenes de datos.

    c) Utiliza algoritmos para buscar soluciones óptimas en grandes espacios de búsqueda, lo que la hace adecuada para problemas de optimización en la RI.

    d) Es útil solo para problemas de clasificación de documentos y no para otras tareas de RI en general.

18. Se tiene un grafo donde cada nodo es un personaje de cierto libro de cuentos y la existencia de las aristas está definida si dos personajes aparecen en el mismo cuento. Se puede asegurar que:

    a) La centralidad de intermediación identifica a los personajes que actúan como conectores entre personajes de cuentos distintos.

    b) La centralidad de grado es útil para identificar los personajes con más conexiones dentro de la red, lo que puede indicar su importancia en el libro.

    c) La centralidad de cercanía ofrece una relación entre la cantidad de vecinos de un nodo con respecto a la longitud máxima de un camino dentro del grafo partiendo del nodo en cuestión.

    d) La centralidad de cercanía indica el grado de conexión de cada personaje con el resto de los personajes de los cuentos del libro.

19. Considerando las dimensiones y desafíos inherentes a Big Data puede afirmarse, tomando en cuenta las características clave y las implicaciones para su procesamiento y análisis, que:

    a) Big Data se caracteriza principalmente por su pequeño volumen y uniformidad, permitiendo un procesamiento eficiente con mínimas adaptaciones de las herramientas de análisis de datos tradicionales.

    b) Big Data no desafía la capacidad de las herramientas tradicionales de procesamiento de datos para capturar, almacenar, gestionar y analizar efectivamente la información, dada la evolución constante de las capacidades computacionales y algoritmos de optimización.

    c) Aunque Big Data puede incluir datos estructurados, su naturaleza se expande al incorporar grandes cantidades de datos no estructurados y semiestructurados, lo que exige el uso de tecnologías especializadas en almacenamiento y procesamiento como Hadoop y sistemas de bases de datos NoSQL.

    d) Además de su complejidad y diversidad, Big Data introduce desafíos significativos en términos de veracidad y variabilidad al requerir métodos avanzados de limpieza y validación de datos para asegurar la integridad del análisis.

20. El algoritmo de PageRank puede describirse como un procedimiento utilizado para:

    a) Determinar la velocidad de carga de una página web en un navegador.

    b) Calcular la relevancia de una página web en función de la cantidad y calidad de los enlaces que apuntan hacia ella.

    c) Prevenir el spam y el contenido no deseado en las páginas web.

    d) Clasificar las páginas web en función de su edad y autoridad.

Trabajo de Control
Sistemas de Recuperación de Información
Fecha: 15 de julio de 2024
Temario #66

Nombre: _____ Grupo: _____

Seleccione en cada caso la(s) respuesta(s) correcta(s). Considere también no marcar ninguna opción.

1. La Web 2.0 se caracteriza por:
   a) La existencia de sitios web dinámicos e interactivos que permiten a los usuarios participar, comentar e interactuar tanto con los creadores de contenido como con otros usuarios.
   b) El uso de distintas tecnologías para crear experiencias web más interactivas y con mayor capacidad de respuesta.
   c) Las actualizaciones del contenido de los sitios web se efectúa de forma manual dentro del código HTML
   d) Las plataformas ofrecen experiencias a la medida, permitiendo a los usuarios personalizar sus perfiles, recibir recomendaciones ajustadas al contenido y participar en filtrado colaborativo.

2. En el contexto del procesamiento de grandes conjuntos de datos, se puede asegurar que:
   a) MapReduce es un enfoque para procesar datos de forma secuencial en un solo servidor para evitar problemas de concurrencia.
   b) MapReduce divide una tarea en múltiples pasos de map y reduce que se ejecutan de forma secuencial en diferentes servidores para mejorar el rendimiento y la escalabilidad.
   c) MapReduce no es adecuado para el procesamiento de datos no estructurados.
   d) MapReduce solo puede manejar pequeñas cantidades de datos y no escala bien a grandes conjuntos de datos.

3. En el contexto de la RI en redes sociales se puede afirmar que:
   a) El análisis de redes sociales se centra exclusivamente en la cantidad de seguidores que tiene un usuario en particular para determinar su influencia en la red.
   b) El análisis de centralidad de intermediación se utiliza para identificar usuarios que son importantes en una red social debido a su posición como "puentes" entre diferentes grupos de usuarios.
   c) El análisis de sentimientos se utiliza para determinar la popularidad de una publicación en redes sociales sin tener en cuenta la opinión de los usuarios.
   d) El análisis de redes sociales no es útil para comprender la difusión de información en una red social específica, para ello se utiliza la medida de centralidad de vector propio.

4. Para contribuir positivamente al posicionamiento orgánico de un sitio web en los motores de búsqueda se puede:
   a) Crear contenido relevante y de alta calidad que satisfaga las necesidades de información de los usuarios.
   b) Mejorar la velocidad de carga del sitio web y asegurar que sea *responsive* y fácil de usar en dispositivos móviles.
   c) Incluir una densidad alta de palabras clave para asegurar que el sitio web aparezca en tantas búsquedas como sea posible.
   d) Obtener enlaces entrantes de otros sitios web de alta autoridad y relevancia temática.

5. La diferencia entre la indexación por tokens y la indexación por conceptos puede definirse como:
   a) La indexación por tokens asigna pesos a los términos basados en su importancia relativa, mientras que la indexación por conceptos utiliza un sistema de etiquetado para asociar términos con características generales.
   b) La indexación por tokens divide los datos en términos individuales, mientras que la indexación por conceptos agrupa los datos en categorías definidas.
   c) La indexación por tokens asigna un valor numérico a cada término de los datos, mientras que la indexación por conceptos utiliza algoritmos de encriptación para proteger la privacidad de los datos.
   d) La indexación por tokens normaliza los datos reduciéndolos a su forma básica, mientras que la indexación por conceptos utiliza un método de ordenación para organizar los términos característicos de los datos.

6. ¿Qué es un "Uniform Resource Locator (URL) Frontier" en el contexto de Web Crawling?
   a) Un protocolo que define cómo se deben formatear las URLs para el crawling.
   b) Una lista prioritaria de URLs que aún no han sido visitadas por el crawler.
   c) Una técnica para filtrar URLs irrelevantes y mejorar la eficiencia del crawling.

d) Una base de datos que almacena URLs únicas identificadas como recursos en la web.

7. La Web 1.0 se caracteriza por:
   a) Las actualizaciones del contenido de los sitios web se efectúa de forma manual dentro del código HTML
   b) Los sitios web se centran en brindar información en lugar de facilitar la colaboración o participación de los usuarios.
   c) Uso de comunidades virtuales para popularizar los sitios web de noticias.
   d) Los propietarios de los sitios web proporcionan contenido de forma periódica.

8. Un investigador necesita recopilar datos de múltiples sitios web para un estudio académico, pero se enfrenta a varios desafíos al realizar el proceso de Web Scraping de manera ética y legal. ¿Cuál de las siguientes opciones describe mejor uno de los desafíos asociados al proceso de Web Scraping?
   a) La necesidad de comprender la estructura del sitio web y su código HTML para extraer los datos correctamente.
   b) La disponibilidad limitada de datos en línea que se pueden extraer utilizando técnicas de Web Scraping.
   c) La dificultad para encontrar herramientas de Web Scraping gratuitas y fiables.
   d) La necesidad de estar montado sobre un scrawler que cumpla con todas las políticas.

9. La premisa básica del algoritmo de PageRank para clasificar páginas web en los resultados de búsqueda es:
   a) La antigüedad de la página web es el principal factor para determinar su clasificación.
   b) El contenido y la relevancia de las palabras clave en la página web determinan su posición.
   c) Los enlaces entrantes a una página web desde otras páginas contribuyen a su importancia y clasificación.
   d) La cantidad de visitas que recibe una página web determina su clasificación.

10. Los Web Crawlers se enfrentan a desafíos constantes. Dentro de ellos se encuentran:
    a) La dificultad para generar el contenido dinámico en tiempo real.
    b) La modificación del código y la estructura del sitio web.
    c) La incapacidad para interpretar correctamente el lenguaje de programación utilizado en el desarrollo de los sitios web.
    d) La falta de acceso a la base de datos del servidor web para extraer información actualizada.

11. El posicionamiento de un sitio web en los motores de búsqueda puede ser afectado por:
    a) Utilizar técnicas de encubrimiento para mostrar contenido diferente a lo indexado por los motores de búsqueda.
    b) Contenido sin valor en el sitio web.
    c) Mantener una estructura de URL clara y coherente.
    d) Obtener enlaces de sitios web irrelevantes y de baja calidad.

12. Al "relajar" el concepto de clique en la detección de comunidades se intenta solucionar:
    a) La imposibilidad de conectar nodos distantes.
    b) La necesidad de datos externos para analizar la red.
    c) La uniformidad de los nodos en términos de grado.
    d) El solapamiento y la complejidad computacional.

13. Dentro del ecosistema de Hadoop, el HDFS se caracteriza por:
    a) El modelo de acceso y de escritura de datos en tiempo real.
    b) La tolerancia a fallos mediante la replicación de datos.
    c) El almacenamiento exclusivo para archivos de texto.
    d) La capacidad ilimitada de almacenamiento.

14. Una plataforma de comercio electrónico desea mejorar la experiencia del usuario al permitir una navegación más personalizada y contextualizada. Actualmente los usuarios tienen dificultades para encontrar productos relevantes debido a la gran cantidad de opciones disponibles. La empresa está interesada en implementar características de la Web 2.5 y la Web Semántica para abordar este problema. ¿Qué características podrían ayudar para ofrecer una navegación más personalizada y contextualizada?
    a) La implementación de ontologías y metadatos para enriquecer la descripción de productos y mejorar la precisión de las recomendaciones.
    b) La optimización de la velocidad de carga del sitio web para mejorar la experiencia del usuario y reducir el abandono del carrito de compra.
    c) La integración de redes sociales para permitir la recomendación de productos basada en las preferencias de amigos y contactos.
    d) La incorporación de los productos en tendencia en el mercado.

15. En un sistema de control de tráfico urbano basado en reglas, ¿cuál de las siguientes reglas sería más efectiva para manejar situaciones de congestión en una intersección?

a) Si hay pocos vehículos en la intersección, reducir el tiempo de los semáforos en verde.

b) Si hay muchos vehículos en la intersección, aumentar el tiempo de los semáforos en verde.

c) Si hay un vehículo de emergencia en la intersección, detener todos los demás vehículos.

d) Si hay muchos peatones cruzando la intersección, reducir el tiempo de los semáforos en rojo.

16. En la RI en el contexto de Big Data, se puede asegurar que:

a) El almacenamiento distribuido centraliza todos los datos en un único servidor para facilitar su gestión y mantenimiento.

b) El almacenamiento distribuido ofrece ventajas significativas en términos de escalabilidad y rendimiento en comparación con el almacenamiento centralizado.

c) El almacenamiento distribuido reparte los datos en múltiples servidores para mejorar la disponibilidad y la redundancia del sistema.

d) El almacenamiento distribuido es una técnica obsoleta.

17. La centralidad de intermediación de un nodo indica:

a) La cantidad de nodos vecinos directos.

b) La resistencia del nodo a fallos.

c) El número total de conexiones entrantes y salientes.

d) La frecuencia con la que un nodo actúa como puente en el camino más corto entre otros dos nodos.

18. Los algoritmos para detectar comunidades en una red intentan:

a) Buscar conjuntos donde cada nodo de un mismo conjunto tenga características similares al resto de los nodos del conjunto.

b) Seleccionar aleatoriamente nodos de alto grado y sus vecinos.

c) Buscar subgrafos tal que no incluyan nodos cuya ausencia desconecte al subgrafo.

d) Encontrar grupos donde los nodos pertenecientes a los mismos grupos son cercanos bajo cierta métrica y lejanos con respecto a los nodos de otros grupos.

19. En el modelo de representación del conocimiento basado en herencia se puede asegurar que:

a) La cantidad de padres que puede tener un nodo no es mayor que 4.

b) Las conclusiones pueden ser canceladas si el grafo es ambiguo.

c) Solo se usa en entornos referentes a la biología.

d) El razonamiento deducido está respaldado por al menos un camino dentro del grafo.

20. El conjunto frontera de URLs en un Web Crawler:

a) Delimita los sitios web que visitará en futuras iteraciones del proceso.

b) Es similar a un conjunto de URLs que esperan ser visitadas.

c) Almacena hipervínculos que pertenecen al mismo dominio del conjunto semilla de URLs con que inició el crawler.

d) Indica las secciones que no pueden ser visitadas de cada sitio web.

Trabajo de Control
Sistemas de Recuperación de Información
Fecha: 15 de julio de 2024
Temario #67

Nombre: _____ Grupo: _____

Seleccione en cada caso la(s) respuesta(s) correcta(s). Considere también no marcar ninguna opción.

1. En la RI en el contexto de Big Data, se puede asegurar que:
   a) El almacenamiento distribuido centraliza todos los datos en un único servidor para facilitar su gestión y mantenimiento.
   b) El almacenamiento distribuido ofrece ventajas significativas en términos de escalabilidad y rendimiento en comparación con el almacenamiento centralizado.
   c) El almacenamiento distribuido reparte los datos en múltiples servidores para mejorar la disponibilidad y la redundancia del sistema.
   d) El almacenamiento distribuido es una técnica obsoleta.

2. El concepto de *rank sink* en el algoritmo de PageRank representa páginas web:
   a) Con distribuciones uniformes de la puntuación de PageRank.
   b) Que tienen una puntuación de PageRank más alta que otras debido a la manipulación de enlaces entrantes y salientes.
   c) Con una baja calidad de contenido y una cantidad insuficiente de enlaces salientes, lo que las hace menos relevantes en los resultados de búsqueda.
   d) Con un alto número de enlaces salientes que no reciben enlaces entrantes, lo que puede afectar negativamente su puntuación de PageRank.

3. Una plataforma de comercio electrónico desea mejorar la experiencia del usuario al permitir una navegación más personalizada y contextualizada. Actualmente los usuarios tienen dificultades para encontrar productos relevantes debido a la gran cantidad de opciones disponibles. La empresa está interesada en implementar características de la Web 2.5 y la Web Semántica para abordar este problema. ¿Qué características podrían ayudar para ofrecer una navegación más personalizada y contextualizada?
   a) La implementación de ontologías y metadatos para enriquecer la descripción de productos y mejorar la precisión de las recomendaciones.
   b) La optimización de la velocidad de carga del sitio web para mejorar la experiencia del usuario y reducir el abandono del carrito de compra.
   c) La integración de redes sociales para permitir la recomendación de productos basada en las preferencias de amigos y contactos.
   d) La incorporación de los productos en tendencia en el mercado.

4. En una empresa comercial, se tiene una red donde los nodos corresponden a los empleados y las aristas representan la frecuencia con la que colaboran en las ventas. Se busca mejorar la colaboración entre los empleados para aumentar las ventas totales, por lo que la directiva debe:
   a) Utilizar la centralidad de cercanía para identificar los empleados menos cercanos a otros en la red y asignarles tareas individuales para evitar posibles conflictos y desacuerdos en el proceso de colaboración.
   b) Utilizar la centralidad de cercanía para identificar los empleados más cercanos a otros en la red y promover la colaboración entre ellos, facilitando así la comunicación y el intercambio de conocimientos para mejorar las ventas.
   c) No considerar la centralidad de cercanía de los empleados al diseñar estrategias de colaboración, ya que todos los empleados tienen el mismo acceso a la información y los recursos.
   d) No tener en cuenta la centralidad de cercanía de los empleados al diseñar estrategias de colaboración, ya que esto podría introducir complicaciones adicionales en el proceso de trabajo.

5. La integración de Hadoop y MapReduce en la RI trae como ventaja:
   a) La eliminación de la necesidad de sistemas de bases de datos.
   b) La posibilidad de la extracción de información relevante y la generación de resultados significativos de grandes conjuntos de datos.
   c) La garantía de la privacidad absoluta de los datos procesados.
   d) La reducción de los costos operativos a cero.

6. Se tiene un grafo donde cada nodo es un personaje de cierto libro de cuentos y la existencia de las aristas está definida si dos personajes aparecen en el mismo cuento. Se puede asegurar que:

a) La centralidad de intermediación identifica a los personajes que actúan como conectores entre personajes de cuentos distintos.

b) La centralidad de grado es útil para identificar los personajes con más conexiones dentro de la red, lo que puede indicar su importancia en el libro.

c) La centralidad de cercanía ofrece una relación entre la cantidad de vecinos de un nodo con respecto a la longitud máxima de un camino dentro del grafo partiendo del nodo en cuestión.

d) La centralidad de cercanía indica el grado de conexión de cada personaje con el resto de los personajes de los cuentos del libro.

7. Si se tiene el conjunto de páginas interconectadas $\{A \to B, C; B \to C; C \to A\}$, entonces la página con valor más alto de PageRank es:

a) A y B.

b) A.

c) C.

d) B.

8. El uso de Hadoop y MapReduce en el contexto de la RI tiene como objetivo:

a) Automatizar el mantenimiento de sistemas de bases de datos.

b) Facilitar el análisis en tiempo real de datos de redes sociales.

c) Mejorar la eficiencia energética en centros de datos.

d) Procesar y analizar grandes conjuntos de datos para la RI.

9. Un sistema cuenta con la siguiente información:

- Juan pertenece a la marina.
- Juan es capellán (sacerdote encargado de una tarea específica fuera de la parroquia).
- Los infantes de la marina suelen ser bebedores de cerveza.
- Un capellán no suele ser bebedor de cerveza.
- Un bebedor de cerveza suele tener sobrepeso.
- Por lo general, un infante de la marina está en buena condición física.

Si se representa la información en una red de herencia se puede concluir que:

a) No se puede asegurar que Juan sea capellán.

b) No se puede asegurar que Juan esté en buena condición física.

c) Juan tiene sobrepeso.

d) Existen dos razonamientos cancelables.

10. ¿Qué estrategia utilizan los Web Crawlers para asegurar un rastreo eficiente y respetuoso de los recursos de los sitios web?

a) Ignorar completamente el archivo `robots.txt` de los sitios web.

b) Visitar y rastrear todos los enlaces de una página web simultáneamente.

c) Extraer únicamente contenido multimedia para reducir la carga en los servidores web.

d) Seguir las directrices del archivo `robots.txt` y aplicar un retraso entre las solicitudes.

11. La afirmación que mejor refleja el principio subyacente de PageRank, considerando su importancia en la clasificación de los sitios web, es:

a) PageRank valora más la cantidad de enlaces entrantes a una página web, independientemente de la calidad o relevancia de estos enlaces.

b) La efectividad de PageRank se basa exclusivamente en el análisis de las palabras clave contenidas en los enlaces entrantes, sin considerar la estructura de enlace de la Web.

c) El algoritmo de PageRank considera tanto la cantidad como la calidad de los enlaces entrantes, asignando mayor valor a los enlaces provenientes de sitios web considerados como "importantes".

d) PageRank opera bajo el supuesto de que los enlaces entrantes y salientes tienen el mismo impacto en la valoración de la relevancia de una página web.

12. Sobre el algoritmo de PageRank, visto en clase, se puede afirmar que:

a) Evalúa la importancia de un sitio web en función de la calidad y cantidad de enlaces entrantes que recibe de otros sitios web.

b) Solo tiene en cuenta el contenido en un sitio web para determinar su relevancia en los resultados de búsqueda.

c) Asigna una puntuación alta a los sitios web que tienen un gran número de enlaces entrantes sin tener en cuenta la calidad de esos enlaces.

d) Asigna una puntuación baja a los sitios web que contienen muchos enlaces salientes, ya que indica una falta de relevancia.

13. El propósito de la política de revisitado en los Web Crawlers es:

a) Determinar la frecuencia con la que los crawlers deben volver a visitar una URL específica para mantener la información actualizada en el índice de búsqueda.

b) Establecer reglas sobre el tiempo máximo que los crawlers pueden pasar en un sitio web durante cada visita para evitar sobrecargar los servidores.

c) Limitar el acceso de los crawlers a ciertos servidores luego de visitar las páginas alojadas en estos.

d) Definir la prioridad de rastreo de las URLs, determinando el orden en que los crawlers visitan y procesan cada página web.

14. Considerando las dimensiones y desafíos inherentes a Big Data puede afirmarse, tomando en cuenta las características clave y las implicaciones para su procesamiento y análisis, que:
    a) Big Data se caracteriza principalmente por su pequeño volumen y uniformidad, permitiendo un procesamiento eficiente con mínimas adaptaciones de las herramientas de análisis de datos tradicionales.
    b) Big Data no desafía la capacidad de las herramientas tradicionales de procesamiento de datos para capturar, almacenar, gestionar y analizar efectivamente la información, dada la evolución constante de las capacidades computacionales y algoritmos de optimización.
    c) Aunque Big Data puede incluir datos estructurados, su naturaleza se expande al incorporar grandes cantidades de datos no estructurados y semiestructurados, lo que exige el uso de tecnologías especializadas en almacenamiento y procesamiento como Hadoop y sistemas de bases de datos NoSQL.
    d) Además de su complejidad y diversidad, Big Data introduce desafíos significativos en términos de veracidad y variabilidad al requerir métodos avanzados de limpieza y validación de datos para asegurar la integridad del análisis.

15. En el contexto de la RI en redes sociales se puede afirmar que:
    a) El análisis de redes sociales se centra exclusivamente en la cantidad de seguidores que tiene un usuario en particular para determinar su influencia en la red.
    b) El análisis de centralidad de intermediación se utiliza para identificar usuarios que son importantes en una red social debido a su posición como "puentes" entre diferentes grupos de usuarios.
    c) El análisis de sentimientos se utiliza para determinar la popularidad de una publicación en redes sociales sin tener en cuenta la opinión de los usuarios.
    d) El análisis de redes sociales no es útil para comprender la difusión de información en una red social específica, para ello se utiliza la medida de centralidad de vector propio.

16. Implementar índices invertidos en un SRI asegura:
    a) Acelerar el proceso de búsqueda al permitir búsquedas directas por contenido en lugar de por título.
    b) Reducir la cantidad de espacio de almacenamiento necesario al comprimir los datos de los documentos.
    c) Facilitar la búsqueda sobre los datos que contienen términos específicos al mantener una lista de datos para cada término único.
    d) Incrementar la seguridad de los datos almacenados al dificultar el acceso directo a la información sin el índice correcto.

17. El algoritmo de PageRank converge si:
    a) La norma de la diferencia entre los vectores es menor a un umbral predefinido.
    b) Finaliza la ejecución cuando el número de iteraciones excede un máximo de iteraciones previamente definido.
    c) Se define un factor de normalización en la fórmula de la función.
    d) El algoritmo no se implementa de forma iterativa.

18. Analizar una red permite:
    a) Detectar posibles tendencias antes de que se conviertan en tendencia.
    b) Obtener predicciones exactas de eventos futuros en mercados financieros.
    c) Evaluar la calidad del contenido de un sitio web.
    d) Encontrar nodos "sensibles" o críticos para la red.

19. La Web 3 se conoce como:
    a) Internet de las cosas.
    b) Web de solo lectura.
    c) Web semántica.
    d) Web de escritura-lectura.

20. La premisa básica del algoritmo de PageRank para clasificar páginas web en los resultados de búsqueda es:
    a) La antigüedad de la página web es el principal factor para determinar su clasificación.
    b) El contenido y la relevancia de las palabras clave en la página web determinan su posición.
    c) Los enlaces entrantes a una página web desde otras páginas contribuyen a su importancia y clasificación.
    d) La cantidad de visitas que recibe una página web determina su clasificación.

Trabajo de Control
Sistemas de Recuperación de Información
Fecha: 15 de julio de 2024
Temario #68

Nombre: _____ Grupo: _____

Seleccione en cada caso la(s) respuesta(s) correcta(s). Considere también no marcar ninguna opción.

1. Implementar índices invertidos en un SRI asegura:
   a) Acelerar el proceso de búsqueda al permitir búsquedas directas por contenido en lugar de por título.
   b) Reducir la cantidad de espacio de almacenamiento necesario al comprimir los datos de los documentos.
   c) Facilitar la búsqueda sobre los datos que contienen términos específicos al mantener una lista de datos para cada término único.
   d) Incrementar la seguridad de los datos almacenados al dificultar el acceso directo a la información sin el índice correcto.

2. El componente responsable de la gestión de recursos y planificación de tareas en Hadoop es:
   a) MapReduce.
   b) Hadoop Common.
   c) HDFS.
   d) YARN.

3. En el algoritmo de Indexación Basada en Clasificación Bloqueada (BSBI), ¿cuál es el paso final para crear un índice invertido para la colección completa de los datos?
   a) Eliminar los términos duplicados de los índices.
   b) Indexar cada bloque de forma independiente.
   c) Fusionar los índices invertidos de cada bloque.
   d) Dividir la colección de datos en bloques de tamaño fijo.

4. Al "relajar" el concepto de clique en la detección de comunidades se intenta solucionar:
   a) La imposibilidad de conectar nodos distantes.
   b) La necesidad de datos externos para analizar la red.
   c) La uniformidad de los nodos en términos de grado.
   d) El solapamiento y la complejidad computacional.

5. Dentro del ecosistema de Hadoop, el HDFS se caracteriza por:
   a) El modelo de acceso y de escritura de datos en tiempo real.
   b) La tolerancia a fallos mediante la replicación de datos.
   c) El almacenamiento exclusivo para archivos de texto.
   d) La capacidad ilimitada de almacenamiento.

6. La Web 2.0 se caracteriza por:
   a) La existencia de sitios web dinámicos e interactivos que permiten a los usuarios participar, comentar e interactuar tanto con los creadores de contenido como con otros usuarios.
   b) El uso de distintas tecnologías para crear experiencias web más interactivas y con mayor capacidad de respuesta.
   c) Las actualizaciones del contenido de los sitios web se efectúa de forma manual dentro del código HTML
   d) Las plataformas ofrecen experiencias a la medida, permitiendo a los usuarios personalizar sus perfiles, recibir recomendaciones ajustadas al contenido y participar en filtrado colaborativo.

7. El algoritmo de PageRank converge si:
   a) La norma de la diferencia entre los vectores es menor a un umbral predefinido.
   b) Finaliza la ejecución cuando el número de iteraciones excede un máximo de iteraciones previamente definido.
   c) Se define un factor de normalización en la fórmula de la función.
   d) El algoritmo no se implementa de forma iterativa.

8. Los algoritmos para detectar comunidades en una red intentan:
   a) Buscar conjuntos donde cada nodo de un mismo conjunto tenga características similares al resto de los nodos del conjunto.
   b) Seleccionar aleatoriamente nodos de alto grado y sus vecinos.
   c) Buscar subgrafos tal que no incluyan nodos cuya ausencia desconecte al subgrafo.
   d) Encontrar grupos donde los nodos pertenecientes a los mismos grupos son cercanos bajo cierta métrica y lejanos con respecto a los nodos de otros grupos.

9. La computación evolutiva:
   a) No es aplicable en la RI debido a la complejidad de los algoritmos evolutivos.
   b) Solo puede manejar conjuntos de datos pequeños y no es escalable a grandes volúmenes de datos.

c) Utiliza algoritmos para buscar soluciones óptimas en grandes espacios de búsqueda, lo que la hace adecuada para problemas de optimización en la RI.

d) Es útil solo para problemas de clasificación de documentos y no para otras tareas de RI en general.

10. La afirmación que mejor refleja el principio subyacente de PageRank, considerando su importancia en la clasificación de los sitios web, es:

a) PageRank valora más la cantidad de enlaces entrantes a una página web, independientemente de la calidad o relevancia de estos enlaces.

b) La efectividad de PageRank se basa exclusivamente en el análisis de las palabras clave contenidas en los enlaces entrantes, sin considerar la estructura de enlace de la Web.

c) El algoritmo de PageRank considera tanto la cantidad como la calidad de los enlaces entrantes, asignando mayor valor a los enlaces provenientes de sitios web considerados como "importantes".

d) PageRank opera bajo el supuesto de que los enlaces entrantes y salientes tienen el mismo impacto en la valoración de la relevancia de una página web.

11. En un sistema donde el conocimiento está orientado a objetos se puede asegurar que:

a) Se enfatiza la atención a la información de la cual se extrajo el conocimiento.

b) Los marcos y las bandas son las estructuras utilizadas para la representación del modelo.

c) No es posible definir especializaciones de los objetos de la vida real dentro del sistema.

d) Las funciones de agregación dificultan poder establecer relaciones entre los objetos.

12. Un SRI es capaz de:

a) Crear los índices asociados a los datos sin tener que analizar cada dato.

b) Generar índices invertidos de manera óptima sin considerar el contexto.

c) Reducir el tiempo de indexación de los datos si utiliza servidores distribuidos dentro de la red para que cada uno ejecute la indexación del mismo conjunto de datos.

d) No necesitar de ningún almacenamiento externo para alojar los índices de los datos.

13. Al realizar la optimización de contenido para SEO debe considerarse:

a) Seleccionar las palabras clave al azar.

b) Utilizar etiquetas de título y meta descripciones únicas y relevantes para cada página.

c) Incluir palabras clave de manera excesiva en el contenido para mejorar el posicionamiento.

d) Crear contenido valioso y original que satisfaga las necesidades de los usuarios.

14. Se puede afirmar que:

a) No es posible identificar subgrupos dentro de una red utilizando análisis de redes.

b) La cantidad de conexiones de un nodo siempre indica su influencia en la red.

c) El tamaño de una red es siempre indicativo de su efectividad en la transmisión de información.

d) Todas las relaciones en una red tienen la misma importancia para el análisis.

15. La Web 1.0 se caracteriza por:

a) Las actualizaciones del contenido de los sitios web se efectúa de forma manual dentro del código HTML

b) Los sitios web se centran en brindar información en lugar de facilitar la colaboración o participación de los usuarios.

c) Uso de comunidades virtuales para popularizar los sitios web de noticias.

d) Los propietarios de los sitios web proporcionan contenido de forma periódica.

16. Analizar una red permite:

a) Detectar posibles tendencias antes de que se conviertan en tendencia.

b) Obtener predicciones exactas de eventos futuros en mercados financieros.

c) Evaluar la calidad del contenido de un sitio web.

d) Encontrar nodos "sensibles" o críticos para la red.

17. El propósito de la política de revisitado en los Web Crawlers es:

a) Determinar la frecuencia con la que los crawlers deben volver a visitar una URL específica para mantener la información actualizada en el índice de búsqueda.

b) Establecer reglas sobre el tiempo máximo que los crawlers pueden pasar en un sitio web durante cada visita para evitar sobrecargar los servidores.

c) Limitar el acceso de los crawlers a ciertos servidores luego de visitar las páginas alojadas en estos.

d) Definir la prioridad de rastreo de las URLs, determinando el orden en que los crawlers visitan y procesan cada página web.

18. El algoritmo de PageRank puede describirse como un procedimiento utilizado para:

a) Determinar la velocidad de carga de una página web en un navegador.

b) Calcular la relevancia de una página web en función de la cantidad y calidad de los enlaces que apuntan hacia ella.

c) Prevenir el spam y el contenido no deseado en las páginas web.

d) Clasificar las páginas web en función de su edad y autoridad.

19. Los Web Crawlers se enfrentan a desafíos constantes. Dentro de ellos se encuentran:

a) La dificultad para generar el contenido dinámico en tiempo real.

b) La modificación del código y la estructura del sitio web.

c) La incapacidad para interpretar correctamente el lenguaje de programación utilizado en el desarrollo de los sitios web.

d) La falta de acceso a la base de datos del servidor web para extraer información actualizada.

20. Sobre WordNet y su aplicación en el procesamiento del lenguaje natural, se puede afirmar que:

a) El diseño de WordNet facilita su integración en aplicaciones multilingües de NLP, aunque su desarrollo original se centró en el inglés.

b) Aunque WordNet es una herramienta valiosa en el NLP, su estructura no incluye información sobre la frecuencia de uso de las palabras en el lenguaje natural.

c) Los synsets facilitan la identificación de relaciones semánticas entre palabras, como la hiperonimia y la meronimia, enriqueciendo tareas de NLP.

d) WordNet proporciona una base para la desambiguación semántica al agrupar palabras con significados similares en synsets.

Trabajo de Control
Sistemas de Recuperación de Información
Fecha: 15 de julio de 2024
Temario #69

Nombre: _____ Grupo: _____

Seleccione en cada caso la(s) respuesta(s) correcta(s). Considere también no marcar ninguna opción.

1. ¿Qué es un "Uniform Resource Locator (URL) Frontier" en el contexto de Web Crawling?
   a) Un protocolo que define cómo se deben formatear las URLs para el crawling.
   b) Una lista prioritaria de URLs que aún no han sido visitadas por el crawler.
   c) Una técnica para filtrar URLs irrelevantes y mejorar la eficiencia del crawling.
   d) Una base de datos que almacena URLs únicas identificadas como recursos en la web.

2. En la RI en el contexto de Big Data, se puede asegurar que:
   a) La indexación distribuida divide los datos en múltiples fragmentos que se almacenan en varios nodos para permitir búsquedas paralelas y mejorar la escalabilidad.
   b) La indexación distribuida no ofrece ventajas en términos de rendimiento y escalabilidad en comparación con la indexación centralizada.
   c) La indexación distribuida es una técnica obsoleta.
   d) La indexación distribuida almacena todos los datos en un solo servidor para facilitar su acceso y búsqueda.

3. El concepto de *rank sink* en el algoritmo de PageRank representa páginas web:
   a) Con distribuciones uniformes de la puntuación de PageRank.
   b) Que tienen una puntuación de PageRank más alta que otras debido a la manipulación de enlaces entrantes y salientes.
   c) Con una baja calidad de contenido y una cantidad insuficiente de enlaces salientes, lo que las hace menos relevantes en los resultados de búsqueda.
   d) Con un alto número de enlaces salientes que no reciben enlaces entrantes, lo que puede afectar negativamente su puntuación de PageRank.

4. Se puede afirmar que:
   a) El término "Big Data" se refiere exclusivamente al volumen de datos que una organización maneja, sin tener en cuenta la velocidad, la variedad y la veracidad de los datos.
   b) Los SRI pueden beneficiarse de MapReduce para mejorar la RI relevante.
   c) MapReduce es un modelo de procesamiento distribuido utilizado para trabajar con grandes volúmenes de datos.
   d) Uno de los desafíos en el procesamiento de Big Data es la capacidad de gestionar y analizar datos provenientes de diversas fuentes y en diferentes formatos de manera eficiente.

5. El uso de Hadoop y MapReduce en el contexto de la RI tiene como objetivo:
   a) Automatizar el mantenimiento de sistemas de bases de datos.
   b) Facilitar el análisis en tiempo real de datos de redes sociales.
   c) Mejorar la eficiencia energética en centros de datos.
   d) Procesar y analizar grandes conjuntos de datos para la RI.

6. Un SRI es capaz de:
   a) Crear los índices asociados a los datos sin tener que analizar cada dato.
   b) Generar índices invertidos de manera óptima sin considerar el contexto.
   c) Reducir el tiempo de indexación de los datos si utiliza servidores distribuidos dentro de la red para que cada uno ejecute la indexación del mismo conjunto de datos.
   d) No necesitar de ningún almacenamiento externo para alojar los índices de los datos.

7. Para transformar el contenido no estructurado de las páginas web en datos estructurados, el Web Scraping:
   a) Extrae información basada en patrones de HTML/CSS.
   b) Analiza los protocolos de red.
   c) Interpreta el código JavaScript en tiempo real.
   d) Convierte de forma automática imágenes a texto.

8. El algoritmo de PageRank converge si:
   a) La norma de la diferencia entre los vectores es menor a un umbral predefinido.
   b) Finaliza la ejecución cuando el número de iteraciones excede un máximo de iteraciones previamente definido.

c) Se define un factor de normalización en la fórmula de la función.

d) El algoritmo no se implementa de forma iterativa.

9. En un SRI la indexación:

a) Consiste en asociar un identificador único a cada dato almacenado en el sistema.

b) Mejora la experiencia del usuario.

c) Permite la organización y la categorización de la información.

d) Facilita la RI relevante.

10. La centralidad de intermediación de un nodo indica:

a) La cantidad de nodos vecinos directos.

b) La resistencia del nodo a fallos.

c) El número total de conexiones entrantes y salientes.

d) La frecuencia con la que un nodo actúa como puente en el camino más corto entre otros dos nodos.

11. El propósito de la política de revisitado en los Web Crawlers es:

a) Determinar la frecuencia con la que los crawlers deben volver a visitar una URL específica para mantener la información actualizada en el índice de búsqueda.

b) Establecer reglas sobre el tiempo máximo que los crawlers pueden pasar en un sitio web durante cada visita para evitar sobrecargar los servidores.

c) Limitar el acceso de los crawlers a ciertos servidores luego de visitar las páginas alojadas en estos.

d) Definir la prioridad de rastreo de las URLs, determinando el orden en que los crawlers visitan y procesan cada página web.

12. En el contexto de la RI en redes sociales se puede afirmar que:

a) El análisis de redes sociales se centra exclusivamente en la cantidad de seguidores que tiene un usuario en particular para determinar su influencia en la red.

b) El análisis de centralidad de intermediación se utiliza para identificar usuarios que son importantes en una red social debido a su posición como "puentes" entre diferentes grupos de usuarios.

c) El análisis de sentimientos se utiliza para determinar la popularidad de una publicación en redes sociales sin tener en cuenta la opinión de los usuarios.

d) El análisis de redes sociales no es útil para comprender la difusión de información en una red social específica, para ello se utiliza la medida de centralidad de vector propio.

13. Sobre el algoritmo de PageRank, visto en clase, se puede afirmar que:

a) Evalúa la importancia de un sitio web en función de la calidad y cantidad de enlaces entrantes que recibe de otros sitios web.

b) Solo tiene en cuenta el contenido en un sitio web para determinar su relevancia en los resultados de búsqueda.

c) Asigna una puntuación alta a los sitios web que tienen un gran número de enlaces entrantes sin tener en cuenta la calidad de esos enlaces.

d) Asigna una puntuación baja a los sitios web que contienen muchos enlaces salientes, ya que indica una falta de relevancia.

14. Los Web Crawlers se enfrentan a desafíos constantes. Dentro de ellos se encuentran:

a) La dificultad para generar el contenido dinámico en tiempo real.

b) La modificación del código y la estructura del sitio web.

c) La incapacidad para interpretar correctamente el lenguaje de programación utilizado en el desarrollo de los sitios web.

d) La falta de acceso a la base de datos del servidor web para extraer información actualizada.

15. Si una red cumple la propiedad de ser un grafo de mundo pequeño, entonces se conoce que:

a) El número de componentes fuertemente conexas está relacionado con la cantidad de grafos K_n presentes.

b) La longitud media del camino entre todo par de vértices es pequeña.

c) El grafo es un anillo regular de grado 5.

d) La red posee pocos vértices.

16. Una empresa de comercio electrónico necesita mejorar su motor de búsqueda para proporcionar resultados más relevantes a sus usuarios. Actualmente, los resultados de la búsqueda no son precisos y los usuarios a menudo encuentran dificultades para encontrar productos específicos. La empresa está considerando implementar una indexación por conceptos para mejorar la relevancia de los resultados de búsqueda. ¿Qué beneficios podría provocar este cambio?

a) Permite adaptarse fácilmente a cambios en el vocabulario y la terminología utilizada en las descripciones de los productos.

b) Ayuda a identificar automáticamente relaciones entre productos, lo que puede mejorar las recomendaciones personalizadas a los usuarios.

c) Facilita la visualización de productos al agruparlos por categorías o características comunes.

d) Mejora la precisión en la búsqueda de productos relacionados, incluso cuando no coinciden exactamente con los términos de búsqueda del usuario.

17. La web actual se enfrenta a problemas como:
    a) Presencia de grandes volúmenes de datos estructurados.
    b) Presencia de una alta calidad en los datos.
    c) Presencia de datos volátiles y distribuidos.
    d) Heterogeneidad en los datos.

18. En el contexto de la representación del conocimiento basada en herencia, ¿qué caracteriza a la herencia cancelable?
    a) La herencia es el resultado del razonamiento no transitivo.
    b) Las conclusiones no están determinadas y dependen del nodo de interés.
    c) Las propiedades heredadas siempre se mantienen y no pueden anularse.
    d) No existe ambigüedad en las conclusiones obtenidas.

19. El algoritmo de PageRank puede describirse como un procedimiento utilizado para:
    a) Determinar la velocidad de carga de una página web en un navegador.
    b) Calcular la relevancia de una página web en función de la cantidad y calidad de los enlaces que apuntan hacia ella.
    c) Prevenir el spam y el contenido no deseado en las páginas web.
    d) Clasificar las páginas web en función de su edad y autoridad.

20. Considerando las dimensiones y desafíos inherentes a Big Data puede afirmarse, tomando en cuenta las características clave y las implicaciones para su procesamiento y análisis, que:
    a) Big Data se caracteriza principalmente por su pequeño volumen y uniformidad, permitiendo un procesamiento eficiente con mínimas adaptaciones de las herramientas de análisis de datos tradicionales.
    b) Big Data no desafía la capacidad de las herramientas tradicionales de procesamiento de datos para capturar, almacenar, gestionar y analizar efectivamente la información, dada la evolución constante de las capacidades computacionales y algoritmos de optimización.
    c) Aunque Big Data puede incluir datos estructurados, su naturaleza se expande al incorporar grandes cantidades de datos no estructurados y semiestructurados, lo que exige el uso de tecnologías especializadas en almacenamiento y procesamiento como Hadoop y sistemas de bases de datos NoSQL.
    d) Además de su complejidad y diversidad, Big Data introduce desafíos significativos en términos de veracidad y variabilidad al requerir métodos avanzados de limpieza y validación de datos para asegurar la integridad del análisis.

Trabajo de Control
Sistemas de Recuperación de Información
Fecha: 15 de julio de 2024
Temario #70

Nombre: _____ Grupo: _____

Seleccione en cada caso la(s) respuesta(s) correcta(s). Considere también no marcar ninguna opción.

1. La afirmación que mejor describe la política de amabilidad en los Web Crawlers es:
   a) Los crawlers se diseñan para acceder a sitios web sin restricciones y extraer datos de manera agresiva para su indexación.
   b) La política de amabilidad de los Web Crawlers dicta que los crawlers deben priorizar ciertos tipos de contenido sobre otros, ignorando completamente ciertas páginas web.
   c) La política de amabilidad establece pautas y reglas sobre cómo los crawlers deben interactuar con los sitios web para minimizar la carga del servidor y respetar las directivas de los administradores del sitio.
   d) Los Web Crawlers son libres de recopilar datos de cualquier sitio web sin restricciones, independientemente de la cantidad de tráfico que generen.

2. En el contexto del procesamiento de grandes conjuntos de datos, se puede asegurar que:
   a) MapReduce es un enfoque para procesar datos de forma secuencial en un solo servidor para evitar problemas de concurrencia.
   b) MapReduce divide una tarea en múltiples pasos de map y reduce que se ejecutan de forma secuencial en diferentes servidores para mejorar el rendimiento y la escalabilidad.
   c) MapReduce no es adecuado para el procesamiento de datos no estructurados.
   d) MapReduce solo puede manejar pequeñas cantidades de datos y no escala bien a grandes conjuntos de datos.

3. El uso de Hadoop y MapReduce en el contexto de la RI tiene como objetivo:
   a) Automatizar el mantenimiento de sistemas de bases de datos.
   b) Facilitar el análisis en tiempo real de datos de redes sociales.
   c) Mejorar la eficiencia energética en centros de datos.
   d) Procesar y analizar grandes conjuntos de datos para la RI.

4. El algoritmo Hypertext Induced Topic Selection (HITS) intenta buscar nodos especiales. Estos son conocidos como:
   a) Autoridades.
   b) Hubs.
   c) Centrales.
   d) Sensibles.

5. Se puede afirmar que:
   a) El término "Big Data" se refiere exclusivamente al volumen de datos que una organización maneja, sin tener en cuenta la velocidad, la variedad y la veracidad de los datos.
   b) Los SRI pueden beneficiarse de MapReduce para mejorar la RI relevante.
   c) MapReduce es un modelo de procesamiento distribuido utilizado para trabajar con grandes volúmenes de datos.
   d) Uno de los desafíos en el procesamiento de Big Data es la capacidad de gestionar y analizar datos provenientes de diversas fuentes y en diferentes formatos de manera eficiente.

6. En un sistema donde el conocimiento está definido a partir de reglas se puede asegurar que:
   a) El nivel de especificidad de las reglas está limitado.
   b) La representación de conocimiento está basada en la lógica proposicional.
   c) El razonamiento se activa en cadena hacia delante.
   d) El orden en que se definen las reglas no altera el razonamiento del sistema.

7. El algoritmo de PageRank puede describirse como un procedimiento utilizado para:
   a) Determinar la velocidad de carga de una página web en un navegador.
   b) Calcular la relevancia de una página web en función de la cantidad y calidad de los enlaces que apuntan hacia ella.
   c) Prevenir el spam y el contenido no deseado en las páginas web.
   d) Clasificar las páginas web en función de su edad y autoridad.

8. La integración de Hadoop y MapReduce en la RI trae como ventaja:
   a) La eliminación de la necesidad de sistemas de bases de datos.

b) La posibilidad de la extracción de información relevante y la generación de resultados significativos de grandes conjuntos de datos.

c) La garantía de la privacidad absoluta de los datos procesados.

d) La reducción de los costos operativos a cero.

9. ¿Cuál de las siguientes opciones describe mejor la diferencia clave entre Web Crawling y Web Scraping?

a) Web Crawling se centra en la exploración y recopilación de enlaces de múltiples sitios web, mientras que Web Scraping se enfoca en la extracción específica de datos de páginas web individuales.

b) Web Crawling se realiza utilizando herramientas de automatización como Selenium WebDriver, mientras que Web Scraping se lleva a cabo mediante el análisis de HTML y CSS.

c) Web Scraping implica el análisis de la estructura y el contenido de las páginas web para extraer datos, mientras que Web Crawling se refiere a la descarga y almacenamiento de páginas web completas.

d) Web Scraping es más eficaz para rastrear e indexar contenido web para motores de búsqueda, mientras que Web Crawling se utiliza principalmente para la extracción de datos en proyectos de investigación.

10. En la RI en el contexto de Big Data, se puede asegurar que:

a) La indexación distribuida divide los datos en múltiples fragmentos que se almacenan en varios nodos para permitir búsquedas paralelas y mejorar la escalabilidad.

b) La indexación distribuida no ofrece ventajas en términos de rendimiento y escalabilidad en comparación con la indexación centralizada.

c) La indexación distribuida es una técnica obsoleta.

d) La indexación distribuida almacena todos los datos en un solo servidor para facilitar su acceso y búsqueda.

11. Analizar una red permite:

a) Detectar posibles tendencias antes de que se conviertan en tendencia.

b) Obtener predicciones exactas de eventos futuros en mercados financieros.

c) Evaluar la calidad del contenido de un sitio web.

d) Encontrar nodos "sensibles" o críticos para la red.

12. La computación evolutiva:

a) No es aplicable en la RI debido a la complejidad de los algoritmos evolutivos.

b) Solo puede manejar conjuntos de datos pequeños y no es escalable a grandes volúmenes de datos.

c) Utiliza algoritmos para buscar soluciones óptimas en grandes espacios de búsqueda, lo que la hace adecuada para problemas de optimización en la RI.

d) Es útil solo para problemas de clasificación de documentos y no para otras tareas de RI en general.

13. La premisa básica del algoritmo de PageRank para clasificar páginas web en los resultados de búsqueda es:

a) La antigüedad de la página web es el principal factor para determinar su clasificación.

b) El contenido y la relevancia de las palabras clave en la página web determinan su posición.

c) Los enlaces entrantes a una página web desde otras páginas contribuyen a su importancia y clasificación.

d) La cantidad de visitas que recibe una página web determina su clasificación.

14. ¿Qué algoritmos permiten obtener información de una red?

a) Índices de centralidad.

b) Detección de comunidades.

c) Hypertext Induced Topic Selection (HITS).

d) PageRank.

15. Se puede afirmar que:

a) No es posible identificar subgrupos dentro de una red utilizando análisis de redes.

b) La cantidad de conexiones de un nodo siempre indica su influencia en la red.

c) El tamaño de una red es siempre indicativo de su efectividad en la transmisión de información.

d) Todas las relaciones en una red tienen la misma importancia para el análisis.

16. Una plataforma de comercio electrónico desea mejorar la experiencia del usuario al permitir una navegación más personalizada y contextualizada. Actualmente los usuarios tienen dificultades para encontrar productos relevantes debido a la gran cantidad de opciones disponibles. La empresa está interesada en implementar características de la Web 2.5 y la Web Semántica para abordar este problema. ¿Qué características podrían ayudar para ofrecer una navegación más personalizada y contextualizada?

a) La implementación de ontologías y metadatos para enriquecer la descripción de productos y mejorar la precisión de las recomendaciones.

b) La optimización de la velocidad de carga del sitio web para mejorar la experiencia del usuario y reducir el abandono del carrito de compra.

c) La integración de redes sociales para permitir la recomendación de productos basada en las preferencias de amigos y contactos.

d) La incorporación de los productos en tendencia en el mercado.

17. La Web 1.0 se caracteriza por:
    a) Las actualizaciones del contenido de los sitios web se efectúa de forma manual dentro del código HTML
    b) Los sitios web se centran en brindar información en lugar de facilitar la colaboración o participación de los usuarios.
    c) Uso de comunidades virtuales para popularizar los sitios web de noticias.
    d) Los propietarios de los sitios web proporcionan contenido de forma periódica.

18. Dentro del análisis de redes, la centralidad de grado mide la importancia de un nodo basándose en:
    a) El grado del nodo.
    b) La cantidad de veces que aparece el nodo en el camino mínimo entre cualquier par de nodos.
    c) La cantidad de vecinos del nodo.
    d) El número de aristas que posee el nodo.

19. El componente responsable de la gestión de recursos y planificación de tareas en Hadoop es:
    a) MapReduce.
    b) Hadoop Common.
    c) HDFS.
    d) YARN.

20. En el algoritmo de Indexación Basada en Clasificación Bloqueada (BSBI), ¿cuál es el paso final para crear un índice invertido para la colección completa de los datos?
    a) Eliminar los términos duplicados de los índices.
    b) Indexar cada bloque de forma independiente.
    c) Fusionar los índices invertidos de cada bloque.
    d) Dividir la colección de datos en bloques de tamaño fijo.

Trabajo de Control
Sistemas de Recuperación de Información
Fecha: 15 de julio de 2024
Temario #71

Nombre: _____ Grupo: _____

Seleccione en cada caso la(s) respuesta(s) correcta(s). Considere también no marcar ninguna opción.

1. Considerando las prácticas éticas y legales en el Web Scraping, se puede asegurar que:
   a) Web Scraping requiere considerar las políticas de `robots.txt` del sitio web objetivo.
   b) Es importante revisar y respetar los términos de servicio del sitio web, así como las leyes aplicables de protección de datos y derechos de autor, antes de realizar Web Scraping.
   c) Web Scraping sobre datos personales sin consentimiento es generalmente aceptado si los datos se utilizan con fines de investigación.
   d) La extracción de datos mediante Web Scraping siempre es legal, independientemente de las leyes locales sobre derechos de autor y privacidad de datos definidos en los sitios web.

2. La afirmación que mejor refleja el principio subyacente de PageRank, considerando su importancia en la clasificación de los sitios web, es:
   a) PageRank valora más la cantidad de enlaces entrantes a una página web, independientemente de la calidad o relevancia de estos enlaces.
   b) La efectividad de PageRank se basa exclusivamente en el análisis de las palabras clave contenidas en los enlaces entrantes, sin considerar la estructura de enlace de la Web.
   c) El algoritmo de PageRank considera tanto la cantidad como la calidad de los enlaces entrantes, asignando mayor valor a los enlaces provenientes de sitios web considerados como "importantes".
   d) PageRank opera bajo el supuesto de que los enlaces entrantes y salientes tienen el mismo impacto en la valoración de la relevancia de una página web.

3. Una buena práctica de SEO para mejorar el posicionamiento de un sitio web en los motores de búsqueda es:
   a) Obtener enlaces de otros sitios web relevantes y de calidad que apunten al sitio.
   b) Copiar contenido directamente de otros sitios web populares para aumentar la cantidad de páginas indexadas.
   c) Utilizar técnicas de encubrimiento para mostrar contenido diferente a lo indexado por los motores de búsqueda.
   d) Llenar el contenido con palabras clave irrelevantes para aumentar la densidad de palabras clave.

4. Considerando las dimensiones y desafíos inherentes a Big Data puede afirmarse, tomando en cuenta las características clave y las implicaciones para su procesamiento y análisis, que:
   a) Big Data se caracteriza principalmente por su pequeño volumen y uniformidad, permitiendo un procesamiento eficiente con mínimas adaptaciones de las herramientas de análisis de datos tradicionales.
   b) Big Data no desafía la capacidad de las herramientas tradicionales de procesamiento de datos para capturar, almacenar, gestionar y analizar efectivamente la información, dada la evolución constante de las capacidades computacionales y algoritmos de optimización.
   c) Aunque Big Data puede incluir datos estructurados, su naturaleza se expande al incorporar grandes cantidades de datos no estructurados y semiestructurados, lo que exige el uso de tecnologías especializadas en almacenamiento y procesamiento como Hadoop y sistemas de bases de datos NoSQL.
   d) Además de su complejidad y diversidad, Big Data introduce desafíos significativos en términos de veracidad y variabilidad al requerir métodos avanzados de limpieza y validación de datos para asegurar la integridad del análisis.

5. Un investigador necesita recopilar datos de múltiples sitios web para un estudio académico, pero se enfrenta a varios desafíos al realizar el proceso de Web Scraping de manera ética y legal. ¿Cuál de las siguientes opciones describe mejor uno de los desafíos asociados al proceso de Web Scraping?
   a) La necesidad de comprender la estructura del sitio web y su código HTML para extraer los datos correctamente.
   b) La disponibilidad limitada de datos en línea que se pueden extraer utilizando técnicas de Web Scraping.
   c) La dificultad para encontrar herramientas de Web Scraping gratuitas y fiables.
   d) La necesidad de estar montado sobre un scrawler que cumpla con todas las políticas.

6. En el contexto de la RI en redes sociales se puede afirmar que:
   a) El análisis de redes sociales se centra exclusivamente en la cantidad de seguidores que tiene un usuario en particular para determinar su influencia en la red.
   b) El análisis de centralidad de intermediación se utiliza para identificar usuarios que son importantes en una red social debido a su posición como "puentes" entre diferentes grupos de usuarios.
   c) El análisis de sentimientos se utiliza para determinar la popularidad de una publicación en redes sociales sin tener en cuenta la opinión de los usuarios.
   d) El análisis de redes sociales no es útil para comprender la difusión de información en una red social específica, para ello se utiliza la medida de centralidad de vector propio.

7. El algoritmo de PageRank puede describirse como un procedimiento utilizado para:
   a) Determinar la velocidad de carga de una página web en un navegador.
   b) Calcular la relevancia de una página web en función de la cantidad y calidad de los enlaces que apuntan hacia ella.
   c) Prevenir el spam y el contenido no deseado en las páginas web.
   d) Clasificar las páginas web en función de su edad y autoridad.

8. ¿Qué es un "Uniform Resource Locator (URL) Frontier" en el contexto de Web Crawling?
   a) Un protocolo que define cómo se deben formatear las URLs para el crawling.
   b) Una lista prioritaria de URLs que aún no han sido visitadas por el crawler.
   c) Una técnica para filtrar URLs irrelevantes y mejorar la eficiencia del crawling.
   d) Una base de datos que almacena URLs únicas identificadas como recursos en la web.

9. La afirmación que mejor describe la política de amabilidad en los Web Crawlers es:
   a) Los crawlers se diseñan para acceder a sitios web sin restricciones y extraer datos de manera agresiva para su indexación.
   b) La política de amabilidad de los Web Crawlers dicta que los crawlers deben priorizar ciertos tipos de contenido sobre otros, ignorando completamente ciertas páginas web.
   c) La política de amabilidad establece pautas y reglas sobre cómo los crawlers deben interactuar con los sitios web para minimizar la carga del servidor y respetar las directivas de los administradores del sitio.
   d) Los Web Crawlers son libres de recopilar datos de cualquier sitio web sin restricciones, independientemente de la cantidad de tráfico que generen.

10. Un sistema cuenta con la siguiente información:
    - Juan pertenece a la marina.
    - Juan es capellán (sacerdote encargado de una tarea específica fuera de la parroquia).
    - Los infantes de la marina suelen ser bebedores de cerveza.
    - Un capellán no suele ser bebedor de cerveza.
    - Un bebedor de cerveza suele tener sobrepeso.
    - Por lo general, un infante de la marina está en buena condición física.

    Si se representa la información en una red de herencia se puede concluir que:
    a) No se puede asegurar que Juan sea capellán.
    b) No se puede asegurar que Juan esté en buena condición física.
    c) Juan tiene sobrepeso.
    d) Existen dos razonamientos cancelables.

11. La premisa básica del algoritmo de PageRank para clasificar páginas web en los resultados de búsqueda es:
    a) La antigüedad de la página web es el principal factor para determinar su clasificación.
    b) El contenido y la relevancia de las palabras clave en la página web determinan su posición.
    c) Los enlaces entrantes a una página web desde otras páginas contribuyen a su importancia y clasificación.
    d) La cantidad de visitas que recibe una página web determina su clasificación.

12. En un sistema de control de tráfico urbano basado en reglas, ¿cuál de las siguientes reglas sería más efectiva para manejar situaciones de congestión en una intersección?
    a) Si hay pocos vehículos en la intersección, reducir el tiempo de los semáforos en verde.
    b) Si hay muchos vehículos en la intersección, aumentar el tiempo de los semáforos en verde.
    c) Si hay un vehículo de emergencia en la intersección, detener todos los demás vehículos.
    d) Si hay muchos peatones cruzando la intersección, reducir el tiempo de los semáforos en rojo.

13. Los algoritmos para detectar comunidades en una red intentan:
    a) Buscar conjuntos donde cada nodo de un mismo conjunto tenga características similares al resto de los nodos del conjunto.

b) Seleccionar aleatoriamente nodos de alto grado y sus vecinos.

c) Buscar subgrafos tal que no incluyan nodos cuya ausencia desconecte al subgrafo.

d) Encontrar grupos donde los nodos pertenecientes a los mismos grupos son cercanos bajo cierta métrica y lejanos con respecto a los nodos de otros grupos.

14. ¿Qué algoritmos permiten obtener información de una red?

a) Índices de centralidad.

b) Detección de comunidades.

c) Hypertext Induced Topic Selection (HITS).

d) PageRank.

15. El componente responsable de la gestión de recursos y planificación de tareas en Hadoop es:

a) MapReduce.

b) Hadoop Common.

c) HDFS.

d) YARN.

16. En el modelo de representación del conocimiento basado en herencia se puede asegurar que:

a) La cantidad de padres que puede tener un nodo no es mayor que 4.

b) Las conclusiones pueden ser canceladas si el grafo es ambiguo.

c) Solo se usa en entornos referentes a la biología.

d) El razonamiento deducido está respaldado por al menos un camino dentro del grafo.

17. La diferencia entre la indexación por tokens y la indexación por conceptos puede definirse como:

a) La indexación por tokens asigna pesos a los términos basados en su importancia relativa, mientras que la indexación por conceptos utiliza un sistema de etiquetado para asociar términos con características generales.

b) La indexación por tokens divide los datos en términos individuales, mientras que la indexación por conceptos agrupa los datos en categorías definidas.

c) La indexación por tokens asigna un valor numérico a cada término de los datos, mientras que la indexación por conceptos utiliza algoritmos de encriptación para proteger la privacidad de los datos.

d) La indexación por tokens normaliza los datos reduciéndolos a su forma básica, mientras que la indexación por conceptos utiliza un método de ordenación para organizar los términos característicos de los datos.

18. No se considera como técnica para detectar comunidades en una red:

a) Analizar la mutualidad de los enlaces.

b) Usar el agrupamiento jerárquico.

c) Utilizar el algoritmo de K-Means.

d) Encontrar cliques de vértices de grado par.

19. En el contexto de la RI en redes, se puede afirmar que:

a) La detección de comunidades no es relevante para la RI en redes, ya que se centra únicamente en la estructura de la red sin considerar el contenido.

b) La detección de comunidades en una red siempre produce resultados objetivos y consistentes independientemente del algoritmo utilizado.

c) La detección de comunidades ayuda a identificar grupos de nodos altamente conectados entre sí, lo que puede ser útil para comprender la estructura y el contenido de la red.

d) La detección de comunidades solo se aplica a redes pequeñas y simples, no a redes grandes y complejas.

20. Un SRI es capaz de:

a) Crear los índices asociados a los datos sin tener que analizar cada dato.

b) Generar índices invertidos de manera óptima sin considerar el contexto.

c) Reducir el tiempo de indexación de los datos si utiliza servidores distribuidos dentro de la red para que cada uno ejecute la indexación del mismo conjunto de datos.

d) No necesitar de ningún almacenamiento externo para alojar los índices de los datos.

Trabajo de Control
Sistemas de Recuperación de Información
Fecha: 15 de julio de 2024
Temario #72

Nombre: _____ Grupo: _____

Seleccione en cada caso la(s) respuesta(s) correcta(s). Considere también no marcar ninguna opción.

1. La Web 1.0 se caracteriza por:
   a) Las actualizaciones del contenido de los sitios web se efectúa de forma manual dentro del código HTML
   b) Los sitios web se centran en brindar información en lugar de facilitar la colaboración o participación de los usuarios.
   c) Uso de comunidades virtuales para popularizar los sitios web de noticias.
   d) Los propietarios de los sitios web proporcionan contenido de forma periódica.

2. Se tiene un grafo donde cada nodo es un personaje de cierto libro de cuentos y la existencia de las aristas está definida si dos personajes aparecen en el mismo cuento. Se puede asegurar que:
   a) La centralidad de intermediación identifica a los personajes que actúan como conectores entre personajes de cuentos distintos.
   b) La centralidad de grado es útil para identificar los personajes con más conexiones dentro de la red, lo que puede indicar su importancia en el libro.
   c) La centralidad de cercanía ofrece una relación entre la cantidad de vecinos de un nodo con respecto a la longitud máxima de un camino dentro del grafo partiendo del nodo en cuestión.
   d) La centralidad de cercanía indica el grado de conexión de cada personaje con el resto de los personajes de los cuentos del libro.

3. En un sistema donde el conocimiento está orientado a objetos se puede asegurar que:
   a) Se enfatiza la atención a la información de la cual se extrajo el conocimiento.
   b) Los marcos y las bandas son las estructuras utilizadas para la representación del modelo.
   c) No es posible definir especializaciones de los objetos de la vida real dentro del sistema.
   d) Las funciones de agregación dificultan poder establecer relaciones entre los objetos.

4. El posicionamiento de un sitio web en los motores de búsqueda puede ser afectado por:
   a) Utilizar técnicas de encubrimiento para mostrar contenido diferente a lo indexado por los motores de búsqueda.
   b) Contenido sin valor en el sitio web.
   c) Mantener una estructura de URL clara y coherente.
   d) Obtener enlaces de sitios web irrelevantes y de baja calidad.

5. ¿Qué estrategia utilizan los Web Crawlers para asegurar un rastreo eficiente y respetuoso de los recursos de los sitios web?
   a) Ignorar completamente el archivo `robots.txt` de los sitios web.
   b) Visitar y rastrear todos los enlaces de una página web simultáneamente.
   c) Extraer únicamente contenido multimedia para reducir la carga en los servidores web.
   d) Seguir las directrices del archivo `robots.txt` y aplicar un retraso entre las solicitudes.

6. ¿Cuál de las siguientes opciones describe mejor la diferencia clave entre Web Crawling y Web Scraping?
   a) Web Crawling se centra en la exploración y recopilación de enlaces de múltiples sitios web, mientras que Web Scraping se enfoca en la extracción específica de datos de páginas web individuales.
   b) Web Crawling se realiza utilizando herramientas de automatización como Selenium WebDriver, mientras que Web Scraping se lleva a cabo mediante el análisis de HTML y CSS.
   c) Web Scraping implica el análisis de la estructura y el contenido de las páginas web para extraer datos, mientras que Web Crawling se refiere a la descarga y almacenamiento de páginas web completas.
   d) Web Scraping es más eficaz para rastrear e indexar contenido web para motores de búsqueda, mientras que Web Crawling se utiliza principalmente para la extracción de datos en proyectos de investigación.

7. La política de ordenación de URLs en los Web Crawlers tiene como aspecto fundamental:
   a) Limitar el acceso de los crawlers a ciertas secciones de un sitio web, evitando el rastreo de URLs consideradas menos importantes o sensibles.
   b) Definir la estructura de la URL de destino, asegurando que estén ordenadas alfabéticamente para facilitar la navegación y la indexación.

c) Establecer la prioridad de rastreo de las URLs, determinando el orden en que los crawlers visitan y procesan cada página web.

d) Determinar la forma en que los crawlers asignan un valor de relevancia a cada URL para clasificarlas en el índice de búsqueda.

8. Se tiene una red de ingredientes donde cada uno representa un nodo y las aristas simbolizan que los ingredientes forman parte de una misma receta. Se busca mejorar la experiencia culinaria mediante la elaboración de combinaciones de ingredientes más interesantes y creativas, para lo cual se debe:

   a) No considerar la centralidad de grado de los ingredientes, ya que todas las combinaciones de ingredientes son igualmente válidas.

   b) Utilizar la centralidad de grado para identificar los ingredientes menos conectados en la red y tomarlos en cuenta para su inclusión en las combinaciones.

   c) No tener en cuenta la centralidad de grado de los ingredientes para no darle mayor importancia a los ingredientes más comunes.

   d) Utilizar la centralidad de grado para identificar los ingredientes más populares en la red y crear combinaciones que incluyan una variedad de ingredientes menos comunes.

9. Un investigador necesita recopilar datos de múltiples sitios web para un estudio académico, pero se enfrenta a varios desafíos al realizar el proceso de Web Scraping de manera ética y legal. ¿Cuál de las siguientes opciones describe mejor uno de los desafíos asociados al proceso de Web Scraping?

   a) La necesidad de comprender la estructura del sitio web y su código HTML para extraer los datos correctamente.

   b) La disponibilidad limitada de datos en línea que se pueden extraer utilizando técnicas de Web Scraping.

   c) La dificultad para encontrar herramientas de Web Scraping gratuitas y fiables.

   d) La necesidad de estar montado sobre un scrawler que cumpla con todas las políticas.

10. El propósito de la política de revisitado en los Web Crawlers es:

    a) Determinar la frecuencia con la que los crawlers deben volver a visitar una URL específica para mantener la información actualizada en el índice de búsqueda.

    b) Establecer reglas sobre el tiempo máximo que los crawlers pueden pasar en un sitio web durante cada visita para evitar sobrecargar los servidores.

    c) Limitar el acceso de los crawlers a ciertos servidores luego de visitar las páginas alojadas en estos.

    d) Definir la prioridad de rastreo de las URLs, determinando el orden en que los crawlers visitan y procesan cada página web.

11. En una empresa comercial, se tiene una red donde los nodos corresponden a los empleados y las aristas representan la frecuencia con la que colaboran en las ventas. Se busca mejorar la colaboración entre los empleados para aumentar las ventas totales, por lo que la directiva debe:

    a) Utilizar la centralidad de cercanía para identificar los empleados menos cercanos a otros en la red y asignarles tareas individuales para evitar posibles conflictos y desacuerdos en el proceso de colaboración.

    b) Utilizar la centralidad de cercanía para identificar los empleados más cercanos a otros en la red y promover la colaboración entre ellos, facilitando así la comunicación y el intercambio de conocimientos para mejorar las ventas.

    c) No considerar la centralidad de cercanía de los empleados al diseñar estrategias de colaboración, ya que todos los empleados tienen el mismo acceso a la información y los recursos.

    d) No tener en cuenta la centralidad de cercanía de los empleados al diseñar estrategias de colaboración, ya que esto podría introducir complicaciones adicionales en el proceso de trabajo.

12. La Web 2.0 se caracteriza por:

    a) La existencia de sitios web dinámicos e interactivos que permiten a los usuarios participar, comentar e interactuar tanto con los creadores de contenido como con otros usuarios.

    b) El uso de distintas tecnologías para crear experiencias web más interactivas y con mayor capacidad de respuesta.

    c) Las actualizaciones del contenido de los sitios web se efectúa de forma manual dentro del código HTML

    d) Las plataformas ofrecen experiencias a la medida, permitiendo a los usuarios personalizar sus perfiles, recibir recomendaciones ajustadas al contenido y participar en filtrado colaborativo.

13. La integración de Hadoop y MapReduce en la RI trae como ventaja:

    a) La eliminación de la necesidad de sistemas de bases de datos.

    b) La posibilidad de la extracción de información relevante y la generación de resultados significativos de grandes conjuntos de datos.

    c) La garantía de la privacidad absoluta de los datos procesados.

    d) La reducción de los costos operativos a cero.

14. Dentro del análisis de redes, la centralidad de grado mide la importancia de un nodo basándose en:
    a) El grado del nodo.
    b) La cantidad de veces que aparece el nodo en el camino mínimo entre cualquier par de nodos.
    c) La cantidad de vecinos del nodo.
    d) El número de aristas que posee el nodo.

15. La afirmación que mejor refleja el principio subyacente de PageRank, considerando su importancia en la clasificación de los sitios web, es:
    a) PageRank valora más la cantidad de enlaces entrantes a una página web, independientemente de la calidad o relevancia de estos enlaces.
    b) La efectividad de PageRank se basa exclusivamente en el análisis de las palabras clave contenidas en los enlaces entrantes, sin considerar la estructura de enlace de la Web.
    c) El algoritmo de PageRank considera tanto la cantidad como la calidad de los enlaces entrantes, asignando mayor valor a los enlaces provenientes de sitios web considerados como "importantes".
    d) PageRank opera bajo el supuesto de que los enlaces entrantes y salientes tienen el mismo impacto en la valoración de la relevancia de una página web.

16. Dentro del ecosistema de Hadoop, el HDFS se caracteriza por:
    a) El modelo de acceso y de escritura de datos en tiempo real.
    b) La tolerancia a fallos mediante la replicación de datos.
    c) El almacenamiento exclusivo para archivos de texto.
    d) La capacidad ilimitada de almacenamiento.

17. En el algoritmo de Indexación Basada en Clasificación Bloqueada (BSBI), ¿cuál es el paso final para crear un índice invertido para la colección completa de los datos?
    a) Eliminar los términos duplicados de los índices.
    b) Indexar cada bloque de forma independiente.
    c) Fusionar los índices invertidos de cada bloque.
    d) Dividir la colección de datos en bloques de tamaño fijo.

18. Sobre el algoritmo de PageRank, visto en clase, se puede afirmar que:
    a) Evalúa la importancia de un sitio web en función de la calidad y cantidad de enlaces entrantes que recibe de otros sitios web.
    b) Solo tiene en cuenta el contenido en un sitio web para determinar su relevancia en los resultados de búsqueda.
    c) Asigna una puntuación alta a los sitios web que tienen un gran número de enlaces entrantes sin tener en cuenta la calidad de esos enlaces.
    d) Asigna una puntuación baja a los sitios web que contienen muchos enlaces salientes, ya que indica una falta de relevancia.

19. La afirmación que mejor describe la política de amabilidad en los Web Crawlers es:
    a) Los crawlers se diseñan para acceder a sitios web sin restricciones y extraer datos de manera agresiva para su indexación.
    b) La política de amabilidad de los Web Crawlers dicta que los crawlers deben priorizar ciertos tipos de contenido sobre otros, ignorando completamente ciertas páginas web.
    c) La política de amabilidad establece pautas y reglas sobre cómo los crawlers deben interactuar con los sitios web para minimizar la carga del servidor y respetar las directivas de los administradores del sitio.
    d) Los Web Crawlers son libres de recopilar datos de cualquier sitio web sin restricciones, independientemente de la cantidad de tráfico que generen.

20. En un grafo una comunidad es:
    a) Un conjunto de nodos que no comparten ninguna similitud estructural o funcional entre sí.
    b) Un conjunto de nodos aislados.
    c) Un conjunto de nodos altamente conectados que forman un subgrafo completamente independiente del resto de la red.
    d) Un conjunto de nodos que están más densamente interconectados entre sí que con los nodos fuera del conjunto.