Capítulo 6

Retorno de la información en la web

Retorno de información en la web

- Las **máquinas de búsqueda en la web** procesan los sitios web y colecciones de documentos.
  - Los documentos son páginas web.
  - Cubren una parte de la web.
  - Se mantiene un **repositorio indexado de páginas web**, usualmente usando índices invertidos.
  - Más aun, se deben actualizar los índices regularmente.
  - Los resultados de las consultas son las páginas web más relevantes para el usuario ordenadas en orden descendente de relevancia.
- Las **máquinas de búsqueda verticales** son personalizadas para **tópicos específicos**; cubren una colección específica de documentos en la web.

Rastreadores web

- **Problema**: En la web la colección de documentos donde hacer las búsquedas no viene dada de antemano.
  - Hay que encontrar y construir la colección
  - **Idea**: Para ello se pueden aprovechar los **hiperenlaces**.
- **Solución**: Los **rastreadores web** son programas que localizan y recolectan información en la web.
  - Se siguen los **hiperenlaces** presentes en documentos conocidos para encontrar otros documentos.
  - Se puede comenzar por un **conjunto semilla** de documentos, por ejemplo:
    - Mapas de sitios, bases de datos previas de documentos.
    - **Páginas populares** (alto tráfico, muchos enlaces entrantes, contenido actualizado frecuentemente).
    - **Sitios de alta autoridad** como: Wikipedia, medios de comunicación, universidades, portales gubernamentales y científicos.
    - **Segmentación temática**: se eligen semillas relevantes a dominios específicos.
  - Toma **meses** realizar un rastreo.

Rastreadores web

- **Análisis interno del contenido**: Un rastreador web primero **explora una página** y para eso extrae:
  - texto visible, metadatos (titulo, descripción, palabras clave), enlaces internos y externos, recursos como imágenes, videos.
- **Detección de contenido duplicado:** Se detecta si es **contenido duplicado**, comparando el contenido con otras páginas ya rastreadas para **decidir prioridad**.
  - Se busca evitar indexar clones o versiones ligeramente modificadas.
- **Filtrado por calidad o spam:** La idea es evaluar si la página debe ser indexada. Se aplican **reglas para descartar páginas.**
  - páginas con poco contenido útil,
  - páginas con exceso de publicidad o enlaces engañosos,
  - paginas que abusan con las palabras clave en el contenido de una página (con el fin de manipular el ranking),
  - sitiios que muestran una versión diferente de la página al rastreador que a los usuarios.
- **Clasificación de la página preliminar**: Se infiere el **tema** o **categoría** de la página,
  - esto ayuda a decidir si se la indexa en ciertos tópicos (p.ej: noticias, salud, comercio, etc.)

Rastreadores web

- **Problema**: Como la web es inmensa, la etapa de rastreo puede tomar demasiado (años o meses).
- **Solución**: El rastreo se hace por varios procesos en varias máquinas ejecutando en paralelo.
  - El conjunto de enlaces a ser rastreados se puede almacenar en una base de datos de enlaces llamada **frontera de rastreo**.
    - Los nuevos enlaces encontrados en las páginas rastreadas se pueden añadir a este conjunto para ser rastreados a continuación.
  - Hay un **coordinador del sistema de rastreo** que contiene la frontera de rastreo.
  - El coordinador hace el **enrutamiento y la asignación de URLs** a procesos de rastreo para balancear la carga de rastreo y optimizar el uso de recursos.
  - Además, el coordinador aplica **políticas de priorización** para no sobrecargar los procesos de rastreo.
    - Estas políticas son para decidir qué enlaces explorar primero, basadas en la importancia del sitio, frecuencia de actualización, y políticas de cortesía.
  - Los rastreadores consultan el coordinador para recibir listas de URLs, y al completar el rastreo reportan los nuevos enlaces detectados para actualizar la frontera de rastreos.
  - En sistemas muy grandes el coordinador puede operar como un servicio distribuido.

Indexado

- Los documentos recolectados por los rastreadores son procesados por un **sistema de indexado**.
- **Problema**: cuando se está haciendo el indexado se debería poder contestar consultas al mismo tiempo.
  - El indexado toma tiempo.
  - Las máquinas de búsqueda en la web no pueden parar.
- **Solución:**
  - El proceso de indexado se ejecuta en varias máquinas.
  - Se crea una nueva copia del índice en lugar de modificar el índice viejo.
  - El índice viejo se usa para contestar consultas.
  - Luego de completar la fase de rastreo y justa antes de iniciar un nuevo indexado, el índice nuevo se convierte en el índice viejo.

Indexado

- Los índices invertidos que sostienen los motores de búsqueda web son colosales en tamaño y complejidad.
  - Se indexan cientos de miles de millones de páginas web, cada una con miles de términos.
  - El índice se actualiza continuamente para reflejar cambios en la web.
- Para cada palabra del índice invertido hay una **lista de ocurrencias en páginas web**. Cada ítem de esta lista contiene típicamente la siguiente información:
  - **docID**: identificador del documento o página web.
  - **Frecuencia** del término en el documento.
  - **Posiciones** donde aparece el término.
  - **Payload** con las siguientes informaciones:
    - **Importancia semántica**: si aparece en el título, encabezado, o metadatos.
    - **Puntajes de relevancia** como TF-IDF.
    - **Contexto sintáctico o semántico**: si aparece en una pregunta, una cita o una lista.
    - **Información para ranking**, como popularidad del documento o autoridad del sitio.

Indexado

- Otra estructura como una **tabla de documentos** se encarga de mapear docID a URL.
  - Esta tabla se consulta al final del proceso, cuando el motor ya ha determinado qué documentos son relevantes.

Búsquedas

- **Problema**: puede haber demasiadas consultas simultáneas para una máquina de búsqueda.
- **Solución**: Usar múltiples máquinas para contestar consultas.
  - Se pueden mantener los índices en memoria
  - Las consultas pueden ser enrutadas a diferentes máquinas para balanceo de carga.

Búsquedas

- Una lista de ocurrencias de un termino en una página web no guarda el **significado del término**.
- La **desambiguación semántica** ocurre en capas superiores del Sistema de búsqueda.
  - El motor analiza **el contexto completo de la búsqueda** del usuario y usa **modelos de lenguaje** para inferir el **sentido más probable del término**.
  - Luego el motor puede **expandir o reformular la consulta** para incluir sinónimos o aclaraciones.
  - Una vez que se recuperan documentos candidatos, se evalúa su **relevancia semántica**.
    - Se **penalizan** los documentos que usan el término en un sentido diferente al buscado.
    - Esto quiere decir que se reduce la relevancia del documento o se lo excluye del conjunto de resultados que se van a mostrar.

Búsquedas

- El índice se divide en **fragmentos** (**shards**) por términos o rangos de documentos.
  - Los **shards por términos** suelen organizarse en función de rangos lexicográficos de términos,
    - aunque también existen estrategias más sofisticadas que agrupan términos por frecuencia de uso o áreas semánticas (p.ej. términos médicos, términos financieros, etc.).
  - Los **shards por rango de documentos** contienen el índice invertido, pero solo para los documentos de su rango.
- Cada shard se aloja en diferentes servidores, optimizados para consultas específicas.
- Cada shard tiene réplicas para tolerancia a fallos y balanceo de carga.
- Se almacenan en **caché** las listas de términos más consultados.

Búsquedas

- Para el **procesamiento de consultas** tenemos:
  - **Análisis de la consulta:**
    - **Tokenización**: divide la consulta en palabras clave.
    - **Normalización**: corrección ortográfica y tipográfica de la consulta, reducir palabras a su raíz (usando técnicas como lematización y stemming), se eliminan stopwords, agrupación de sinónimos y variantes, se hace reconocimiento de entidades.
    - **Desambiguación semántica**: interpreta el sentido de cada término según el contexto.
    - **Expansión**: puede añadir sinónimos o reformular la consulta para mayor eficiencia y precisión.
  - **Recuperación de documentos:**
    - Se consulta el índice invertido para encontrar documentos que contengan los términos relevantes.
    - Se aplican filtros como penalizaciones por baja calidad (duplicados, spam, etc.).
  - **Ranking de resultados:**
    - Cada documento candidato recibe un ranking de relevancia.
    - Se seleccionan los documentos con mayor puntaje.
    - Se ordenan y presentan al usuario con enlaces, títulos y descripciones.

Búsquedas

- Ahora explicamos el **acceso distribuido al índice invertido**.
- **Enrutamiento de la consulta**: cuando llega una consulta, el sistema determina qué shards deben ser consultados.
- **Búsqueda local en cada shard**: cada shard accede a su porción del índice invertido, recupera los documentos que contienen los términos de la consulta y calcula un **ranking local** de relevancia para esos documentos.
- **Fusión y ranking global**: Los resultados de todos los shards se envían a un **coordinador** (**nodo maestro**).
  - Este nodo fusiona las listas y recalcula un ranking global.

Organización de nodos

- Los motores de búsqueda web modernos están organizados como **sistema distribuidos** de múltiples capas
  - donde distintos tipos de nodos cumplen funciones especializadas en el procesamiento de consultas.
- **Tipos de nodos:**
  - Nodos de análisis de consulta.
  - Nodos de recuperación de documentos (index shard nodes)
  - Nodos de ranking global (ranking coordinator nodes)
  - Nodos de presentación (front-end serving nodes): entregan los resultados al usuario.
  - Nodos de indexado.
  - Nodos de rastreo.

Relevancia de documentos en la web

- Las **máquinas de búsqueda en la web tempranas** ordenaban las respuestas basándose solo en relevancia TF-IDF.
- **Pero usar este enfoque para la web tiene sus problemas:**
  - El número de documentos relevantes a una consulta puede ser enorme si solo frecuencia de términos es usada.
  - Muchas páginas que los usuarios quieren ver pueden tener frecuencia de términos baja y no van a obtener valor TF-IDF alto.
  - Puede haber **páginas spam**: Es solo agregar ciertas palabras en una página sin valor para que aparezca en las búsquedas.
  - La mayoría de los usuarios se interesa por las páginas de los **sitios populares**.

Relevancia de documentos en la web

- **Solución**: usar la **popularidad de un sitio web** (por ejemplo, cuánta gente lo visita) para dar rango a sus páginas web en el resultado de las consultas.
  - Pero es difícil encontrar la popularidad real de un sitio.
  - Porque para obtener esta información es necesaria la cooperación del sitio.
  - Algunos sitios podrían mentir sobre esto para obtener un rango mayor.
- **Solución más refinada**: medidas tradicionales de relevancia de una página como TF-IDF pueden ser combinadas con la popularidad del sitio de la página
  - para obtener una **medición global de la relevancia** de la página para una consulta.
  - Las páginas con mayor valor de relevancia pueden retornarse como las respuestas top de la consulta.

Popularidad de sitios web

- Ahora vemos como medir la popularidad de un sitio web.
- **Idea 1**: usar los **archivos de marcadores** de cada usuario (para sus navegadores) para saber qué páginas son populares:
  - Los sitios que aparecen en una gran cantidad de archivos de marcadores se los puede considerar como muy populares.
  - Sin embargo, los archivos de marcadores usualmente son almacenados privadamente y no son accesibles en la web.

Popularidad de sitios web

- Las páginas web tienen información muy importante de la cual carece el texto plano: **hiperenlaces**.
- **Idea 2**: usar el número de enlaces a un sitio *S* como una medida de la popularidad o prestigio del sitio.
  - Contar un enlace a *S* desde cada sitio que lo enlaza a *S*.
  - La popularidad es para sitios, no para páginas web (la mayoría de los enlaces son a la página principal del sitio).
- **Refinamiento de idea 2**: Cuando se computa el prestigio basado en enlaces a un sitio,
  - se le puede dar más peso a los enlaces de sitios que tienen mayor prestigio.

Popularidad de sitios web

- **Idea 3:** Las máquinas de búsqueda pueden llevar la pista de qué fracción de veces los usuarios seleccionan una página retornada como resultado de una búsqueda.
  - Esta medida puede ser usada como una medición de popularidad del sitio.

PageRank

- **Algoritmo de ordenamiento PageRank**
  - Es un algoritmo definido por **Google** como una **medida de popularidad** de una página basada en la popularidad de las páginas que la enlazan.
  - Analiza los enlaces hacia fuera y los enlaces hacia dentro.
  - Se considera a las páginas altamente enlazadas por otras páginas como más importantes (con mayor autoridad) que las páginas con menos enlaces hacia ellas.
  - Se dice que una página *P* tiene un **rango alto** si la suma de los rangos de las páginas que apuntan a *P* tiene un valor alto.

PageRank

- **Algoritmo de ordenamiento PageRank (continuación)**
  - Supongamos que una persona navegando en la web realiza una caminata aleatoria de páginas web de la siguiente manera:
    - El primer paso comienza en una **página web aleatoria**.
    - En los pasos siguientes la persona hace una de las siguientes:
      - Con una probabilidad δ la persona salta a una página web elegida aleatoriamente.,
      - Con una probabilidad $1 - \delta$ la persona elige aleatoriamente uno de los enlaces hacia afuera de la página actual y sigue ese enlace.

PageRank

- **Algoritmo de ordenamiento PageRank (continuación)**
  - Asumiendo la caminata aleatoria anterior, el **PageRank de una página** web es la probabilidad de que esa persona visite la página en un determinado punto del tiempo.
    - Las páginas apuntadas por muchas páginas web es más probable que sean visitadas y van a tener un PageRank más alto.
    - Las páginas apuntadas por páginas web con alto PageRank van a tener una mayor probabilidad de ser visitadas y entonces van a tener un mayor PageRank.

PageRank

- **Algoritmo de ordenamiento PageRank (continuación)**
  - PageRank puede definirse por un **conjunto de ecuaciones lineales**.
  - Primero se les da **identificadores enteros** a las páginas web.
  - La **matriz de probabilidad de salto** T: T[i, j] es la probabilidad que un caminante aleatorio que sigue un enlace afuera de la página *i* siga el enlace hacia la página *j*.
  - Suponiendo que cada enlace de *i* tiene la misma probabilidad de ser seguido, $T[i, j] = 1/N_i$, donde N_i es el número de enlaces afuera de la página *i*.
  - Pero los motores de búsqueda modernos han evolucionado más allá de esta versión uniforme para calcular T.
    - Se puede usar el texto del áncora de un enlace para inferir relevancia del enlace.
    - O enlaces con texto más descriptivo pueden tener mayor peso.

PageRank

- **Algoritmo de ordenamiento PageRank (continuación)**
  - El PageRank de la página *j* puede definirse como:

$$P[j] = \delta/N + (1 - \delta) * \sum_{i=1}^{N} (T[i, j] * P[i])$$

  - Donde δ es una constante entre 0 y 1(usualmente 0,15) y N es el número de páginas.
  - δ representa la probabilidad de un paso en la caminata aleatoria sea un salto aleatorio.
  - El primer término representa elegir la pagina j aleatoriamente, y eso tiene el mismo peso para todas las páginas.
  - El segundo término representa navegar a página j desde otra página.

PageRank

- **Algoritmo de ordenamiento PageRank (continuación)**
  - El **conjunto de ecuaciones** se resuelve usando una técnica iterativa.
  - Se comienza inicializando cada *P*[i] a 1/*N*.
  - Cada paso de la iteración computa nuevos valores para *P*[i] usando los valores de *P* de la iteración previa.
  - La iteración acaba cuando el valor máximo de cambio en un valor de *P*[i] en la iteración va por debajo de un cierto valor de corte.
  - (Max {|P(k+1)[i] – P(k)[i]|: 1 ≤i ≤ N}) < ε
  - Por ejemplo, ε puede ser 10^(-6).

Popularidad de sitios web

- **Problema**: PageRank asigna una medida de popularidad que no considera los términos de la consulta.
- **Solución**: usar las palabras clave en el texto del áncora de los enlaces a una página para juzgar para qué tópicos la página es altamente relevante.
  - La *popularidad basada en texto de áncoras* puede ser usada en combinación con otras medidas de popularidad y con TF-IDF para obtener un ranking de los resultados de una consulta.
- **Implementación 1**: Si se considera el texto de esas áncoras como parte de la página apuntada, entonces TF-IDF toma texto de áncoras en cuenta.

Popularidad de sitios web

- **Implementación 2:** se computa una **medida de popularidad** usando solo páginas que contienen los términos de la consulta en lugar de computar popularidad usando todas las páginas web disponibles.
  - Este enfoque es más costoso porque el cómputo del ranking de popularidad debe ser hecho dinámicamente cuando se recibe una consulta,
  - mientras que PageRank se computa estáticamente una vez y se reutiliza para todas las consultas.
  - El algoritmo Hits se basa en esta idea de implementación.

Relevancia de documentos en la web

- Los valores TF-IDF de una página deben ser combinados con el ranking de popularidad (p.ej. PageRank).
- **Problema**: ¿cómo combinar TF-IDF con otras medidas de popularidad?
  - Este es un secreto mantenido por muchas empresas.
  - Esto ayuda a protegerse de la competencia y de los que quieren producir páginas spam.
- **Idea de solución**: una fórmula que combina puntajes es fija y toma como **parámetros** pesos para cada factor considerado.
  - Se usa un **conjunto de entrenamiento** de resultados de consultas cuyos rangos son fijados por humanos.
  - Usando el mismo, un **algoritmo de entrenamiento automático** puede calcular valores para esos parámetros.

Buscador de Google

- **Características principales del buscador de Google para la web:**
  - **Rastreo**: Google utiliza arañas web (Googlebot) que recorren la web para encontrar nuevas y actualizadas páginas.
  - **Indexación**: Las páginas encontradas son almacenadas en el índice de Google, una base de datos masiva que contiene la copia de todas las páginas web visitadas por los rastreadores.
  - **Algoritmos de búsqueda**: Google emplea complejos algoritmos para interpretar las consultas de los usuarios y encontrar los resultados más relevantes.
    - Estos algoritmos consideran cientos de factores, incluyendo palabras clave, sinónimos y la ubicación del usuario.
  - **Relevancia de respuestas**: Los algoritmos de Google no sólo encuentran coincidencias de palabras clave, sino que también intentan entender el contexto y la intención de la búsqueda para ofrecer los resultados más útiles y precisos.

Buscador de Google

- **Características del buscador de Google (Cont):**
  - **PageRank**: Google clasifica los resultados de búsqueda basándose en la autoridad y relevancia de las páginas, utilizando medidas como el Page-Rank, que evalúa la cantidad y calidad de los enlaces hacia una página.
  - **Personalización**: Las búsquedas se personalizan en función de la historia del usuario, su ubicación y otros datos disponibles.
  - **Calidad del contenido**: Google valora contenido original, de alta calidad y relevante, penalizando a sitios con prácticas de spam o contenido duplicado.
  - **Lenguaje de consulta**: ya hablamos de ello en el archivo de filminas anterior.

Buscador de Google

- **Datos que contiene un enlace devuelto por el buscador de Google:**
  1. **Título de la página**: El encabezado principal, generalmente en azul, que resume de qué trata la página.
  2. **URL**: La dirección web del sitio, a menudo en verde. Esto te da una idea de la fuente del contenido.
  3. **Extracto o snippet**: Un breve fragmento de texto que Google extrae de la página para mostrar una vista previa del contenido. Esto suele estar en gris y puede incluir las palabras clave de tu búsqueda resaltadas en negrita.
  4. **Breadcrumb**: Una ruta de navegación que muestra dónde se encuentra la página dentro de la jerarquía del sitio web.
  5. **Fecha**: En algunos casos, especialmente en artículos y noticias, Google incluye la fecha de publicación o de la última actualización.

Buscador de Google

- El buscador de Google ofrece mucho más que solo enlaces como resultado de las búsquedas:
  1. **Fragmentos destacados**: Resúmenes directos de respuestas que se muestran en la parte superior de la página de resultados, a menudo en formato de párrafo (p.ej. definiciones, descripciones), lista (p.ej. pasos de instrucciones o listas de elementos) o tabla (p.ej. datos comparativos en forma tabular).
  2. **Panel de conocimiento**: Información detallada sobre un tema, persona, lugar o cosa, normalmente a la derecha de los resultados.
  3. **Videos**: Recomendaciones de videos de plataformas como YouTube que son relevant es para tu consulta.
  4. **Imágenes**: Una selección de imágenes relacionadas con tu búsqueda.
  5. **Preguntas relacionadas**: Una lista de preguntas adicionales que otros usuarios han hecho sobre el mismo tema o uno similar, junto con respuestas breves que se pueden expandir para obtener más información.
  6. **Gráficos y estadísticas**: En algunos casos, como para datos financieros o demográficos, Google puede mostrar gráficos interactivos.