Recuperación de la Información en la Web

Índice

- Introducción
- Estructura de la web
- Arquitectura del sistema
- Ranking
- Web spam
- Gestión de datos web
- Conclusiones

Introducción

Introducción

- La web fue concebida por Tim Berners-Lee en 1989
  - Comprobada en diciembre de 1990
- El primer servidor web se lanzó en 1991
- No se preveía el gran auge e importancia que tendría la web en el futuro
- La web ha hecho crecer de manera exponencial el volumen de datos e información
- Gran parte de las tareas actuales no pueden concebirse sin el uso de la web
  - Comercio electrónico
  - Comunicación
  - Banca
  - Entretenimiento
  - Acceso a información

Introducción

- La web es inmensa
  - La información textual se estima en ordenes de petabytes
  - Según http://www.internetlivestats.com/ el número total de webs es superior a 1100 millones
  - Además, en la web se encuentra otro tipo de datos: audio, video, imágenes, documentos

Introducción

- La web es un repositorio de datos extremadamente grande, publico, ubicuo y sin estructura
- Es necesario herramientas eficientes para recuperar, gestionar y filtrar la información de la web
- El gran tamaño de la web, y la gran velocidad a la que cambia, hacen que los sistemas RI para la web sean complejos
- La mayor parte de la información indexada en los sistemas RI es textual
  - No se puede despreciar la información no-textual

Introducción

- Retos relacionados con los datos de los sistemas RI en la web:
  - Datos distribuidos
    - La información se encuentra alojada en un gran número de ordenadores diferentes
    - Los ordenadores están conectados sin ningún tipo de topología (característica de Internet)
  - Alto porcentaje de datos volátiles
    - Los datos se pueden añadir y eliminar fácilmente
    - Estimación: el 50% de la web cambia en unos pocos meses
  - Gran volumen de datos
    - Crecimiento exponencial de la web

Introducción

- Datos sin estructura y redundantes
  - Los documentos HTML no se consideran estructurados, sino semi-estructurados
  - La información puede estar disponible en varias webs
- Calidad de los datos
  - La web es un nuevo medio de comunicación, pero sin proceso editorial
  - Los datos pueden carecer de calidad: inciertos, erróneos, obsoletos, inválidos, pobremente escritos, errores gramaticales, errores ortográficos, errores mal intencionados, etc.
  - Ejemplo: los errores en los nombres en lenguas extranjeras suelen abundar
- Heterogeneidad de los datos
  - Diferentes tipos de medios/formatos
  - Diferentes lenguajes/alfabetos

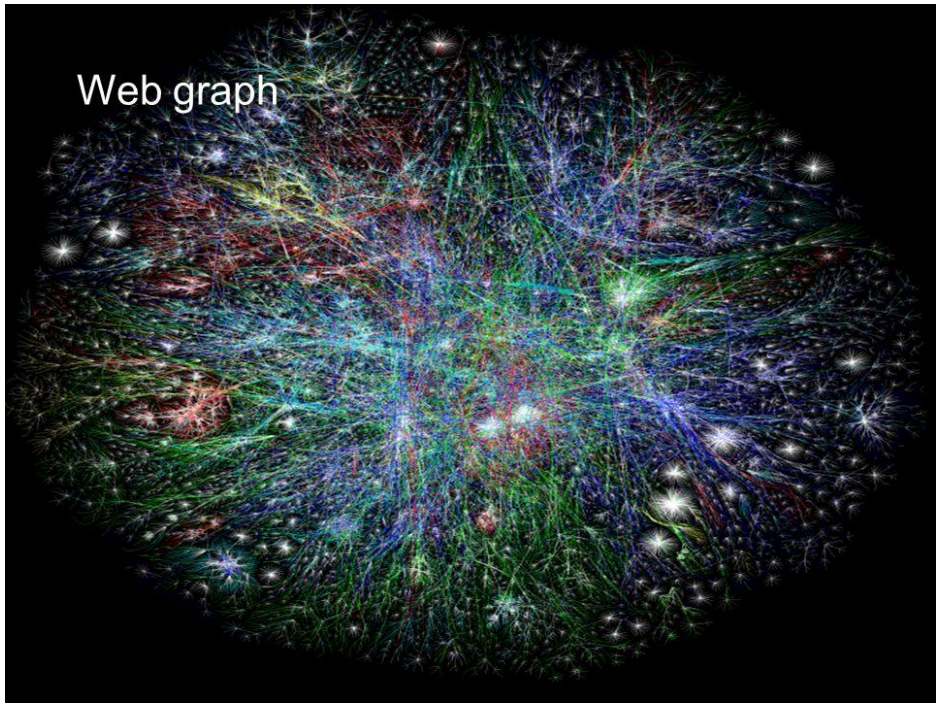
Introducción

- Retos relacionados con los usuarios de los sistemas RI en la web:
  - Expresión de las consultas
    - Las necesidades de los usuarios pueden ser complejas, por lo que no siempre pueden expresarse en una consulta
    - Las necesidades, aún expresadas en lenguaje natural, son imperfectas
  - Interpretación de los resultados
    - Aunque la consulta se haya expresado perfectamente
      - Millones de resultados
      - Ningún resultado
- En resumen
  - El reto para el usuario es expresar adecuadamente la consulta
  - El reto para el sistema RI es realizar una búsqueda rápida que devuelva documentos relevantes, aún cuando la consulta esté pobremente expresada

Estructura de la Web

Estructura de la Web

- La web puede interpretarse como un grafo
  - Los nodos son las páginas web
  - Los vértices son los enlaces entre las páginas



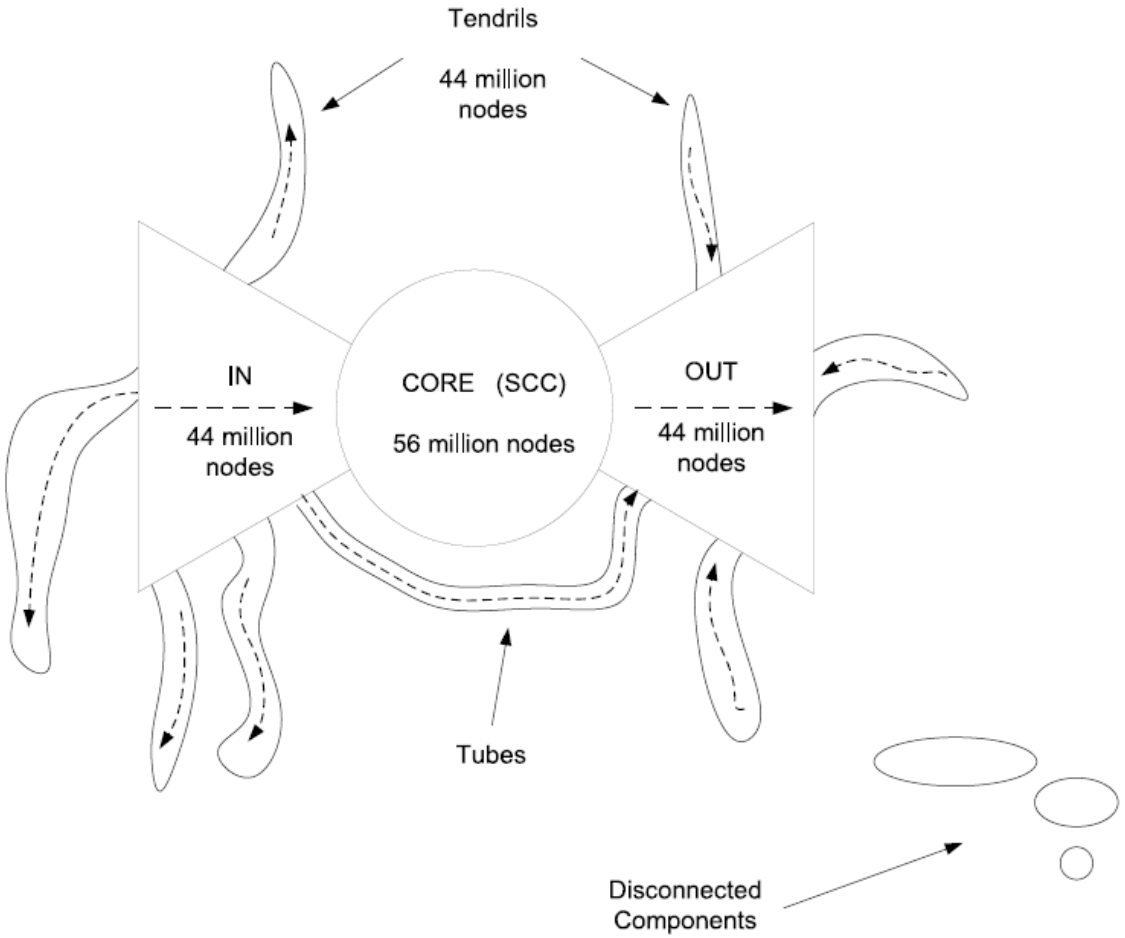
Estructura de la Web

- El grafo de la web puede estructurarse en los siguientes componentes:
  - Núcleo: compuesto por los componentes altamente conectados
    - Se puede navegar desde cualquier punto del núcleo a otro punto aleatorio dentro del núcleo
  - Entrada: puntos desde los que se puede llegar a puntos del núcleo, pero no pueden llegarse desde el núcleo
  - Salida: puntos que pueden alcanzarse desde el núcleo, pero no pueden alcanzar puntos en el núcleo
  - Tubos: puntos que permiten conectar la entrada con la salida sin pasar por el núcleo

Estructura de la Web

- Tentáculos:
  - De entrada: puntos que pueden alcanzarse desde la entrada
  - De salida: puntos que enlazan a puntos en la salida
  - Los tentáculos son puntos que no pertenecen a los otros componentes
- Islas: puntos desconectados de la estructura anterior
  - Interiormente pueden estructurarse de la misma forma

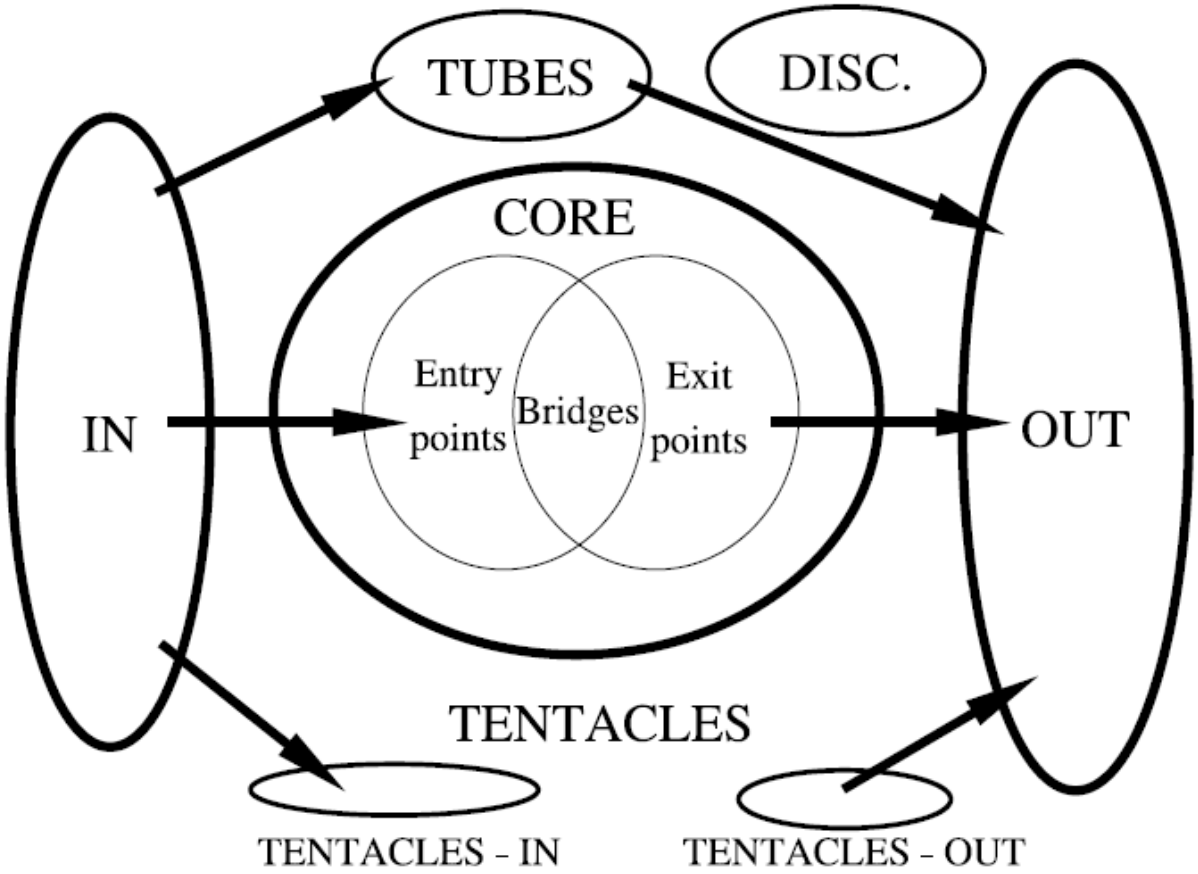
Estructura de la Web



Estructura de la Web

- El núcleo, puede dividirse en:
  - Puentes: sitios que pueden alcanzar directamente las componentes de salida y de entrada
  - Puntos de entrada: sitios que pueden alcanzar la componente de entrada pero no de salida y no son puentes
  - Puntos de salida: sitios que pueden alcanzar la componente de salida directamente, pero no son puentes
  - Normal: no pertenecen a las categorías anteriores

Estructura de la Web



Fuente: Modern Information Retrieval 2nd Edition. Ricardo Baeza-Yates

Arquitectura del Sistema

Arquitectura del Sistema

BÁSICA

Arquitectura del Sistema Básica

- Arquitectura basada en una araña web (crawler) e indexador centralizados
- La araña lanza peticiones a los servidores web remotos para acceder a su información
  - Descarga el texto y los metadatos
- Índice basado en el índice invertido o en alguna de sus variantes
  - Lista de términos, en donde a cada término se le asocia una lista con los documentos en los que aparece
- Se indexa una vista lógica del documento
  - El texto se preprocesa
  - Las palabras vacías se eliminan
    - La elección de palabras vacías se hace estadísticamente

Arquitectura del Sistema Básica

- Para una consulta dada, sólo se muestra un subconjunto de respuestas
  - Los 10 primeros documentos
- Si el usuario necesita más resultados, el sistema recalculará el siguiente subconjunto
- El sistema nunca computara la consulta completa contra todos los documentos
  - Miles de millones de documentos ➔ lento
  - Encontrar unos pocos miles de documentos es suficiente
- El principal problema de la arquitectura básica es que no es capaz de lidiar con la gran cantidad de información que alberga la web

Arquitectura del Sistema

BASADA EN CLUSTER

Arquitectura del Sistema Basada en cluster

- Los sistemas actuales usan una arquitectura masiva y paralela basada en cluster

Arquitectura del Sistema Basada en cluster

- Debido a la gran cantidad de documentos, el índice no cabe en una sola máquina
  - Se tiene que distribuir a lo largo de diversos ordenadores o clusters
  - Los documentos se tienen que dividir en subconjuntos
- La gran cantidad de consultas a las que se enfrenta un sistema RI no puede atenderse desde un único ordenador
  - Replicar la estructura básica para tener un conjunto de clusters que actúen como subsistemas RI
  - Los cluster deben alojarse a lo largo del mundo para evitar latencias
  - La replicación permite la tolerancia a fallos

Arquitectura del Sistema Basada en cluster

- Aspectos a tener en cuenta
  - Lograr un buen balanceo entre las actividades internas (indexación y respuesta) y externas (crawler) del motor RI
    - Clusters dedicados: crawling, indexación, interacción con el usuario, procesamiento de la consulta, generación de la página de resultados, etc.
  - Balancear la carga entre los clusters
  - Preveer el fallo de dispositivos hardware
    - Enviar las consultas a las CPUs disponibles más adecuadas
    - Reemplazar de forma rutinaria discos duros antes de su fallo
    - Uso de componentes hardware intercambiables de bajo coste

Arquitectura del Sistema

CACHÉ

Arquitectura del Sistema Caché

- Un buscador tiene que ser tan rápido como sea posible
  - Siempre que se pueda, las tareas deberían ejecutarse en memoria principal
- Los sistemas de caché son muy recomendable y comúnmente utilizados
  - Permiten tiempos medios de respuesta más cortos
  - Reduce significativamente la carga de trabajo de los servidores back-end
  - Disminuyen el ancho de banda utilizado
- La técnica más efectiva de caché en motores RI es *caching answer* realizada en el front-end
  - Se almacenan los resultados de las consultas más frecuentes
  - Las consultas siguen una distribución de poder: un pequeño grupo de preguntas se repite muchas veces

Arquitectura del Sistema Caché

- En cualquier ventana temporal una gran fracción de consultas serán únicas
  - No existirá caché para ellas.
- El rendimiento puede mejorar incluyendo una caché de listas invertidas del índice a nivel de cluster de búsqueda
- El balance entre los dos tipos de caché no es trivial
  - Analizar los logs de las consultas para extraer conocimiento
- La caché se puede organizar:
  - Resultados precalculados obtenidos de consultas anteriores
  - Listas invertidas de los términos más frecuentes de las consultas

Ranking

Ranking

- El ranking es la tarea principal de cualquier sistema RI
- Los documentos en la web pueden recuperarse siguiendo un modelo tradicional
  - Modelo espacio vectorial
- La web puede representarse como un grafo, donde las webs están relacionadas entre sí.
- Los modelos de RI clásicos pueden mejorarse para tener en cuenta la estructura de red
- Los documentos web, tienen cierta estructura y contenido al que puede darse un mayor peso
- Ciertos dominios (.edu) pueden tener un mayor prestigio
  - Mayor peso en el índice

Ranking

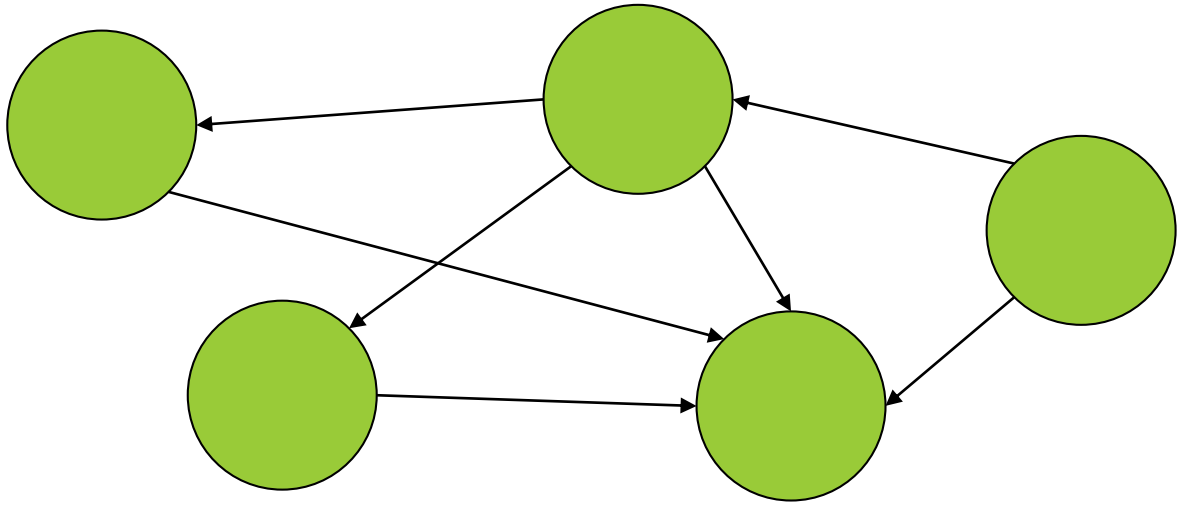
- Tipos de señales utilizadas para mejorar el ranking
  - Señales de contenido: relacionadas con el texto en sí
    - Conteo de palabras
    - Disposición del documento HMLT: título, cabeceras (h1, h2, etc.), tamaños, tipos de fuentes, resaltados, metadatos, etc.
      - Algunas partes pueden tener un mayor peso
    - Proximidad de ciertas etiquetas en la página
  - Señales de estructura: relacionada con la estructura de red de la web
    - Anchor text: describe el contenido de la web a la que apunta
    - Número de enlaces salientes o entrantes

Ranking

- Señales de uso: clics
  - Número de clics en las URLs de los resultados
  - Información del usuario
    - Contexto geográfico (IP, lenguaje)
    - Contexto tecnológico (sistema operativo, navegador, dispositivo móvil)
    - Contexto temporal: histórico de consultas (es necesario el uso de cookies o el registro por parte del usuario)

Ranking

- El número de enlaces que apuntan a una web nos ofrece una medida de su popularidad y calidad
- Relaciones entre páginas importantes para el ranking
  - Enlaces en común en varias páginas
  - Páginas referenciadas por una misma página



Ranking

- Los rankings basados en web miden la importancia de una web dentro de la colección
  - El texto dentro de una web se indexará utilizando un índice invertido y una ponderación de los términos (TF-IDF)
  - Para una consulta dada, el sistema recuperará las web que contengan los términos de la consulta (mediante el índice) y posteriormente ordenara los resultados mediante la importancia de dicha web
- El proceso mejora si integramos la importancia de la web con los modelos clásicos de RI (modelo vectorial o probabilístico)

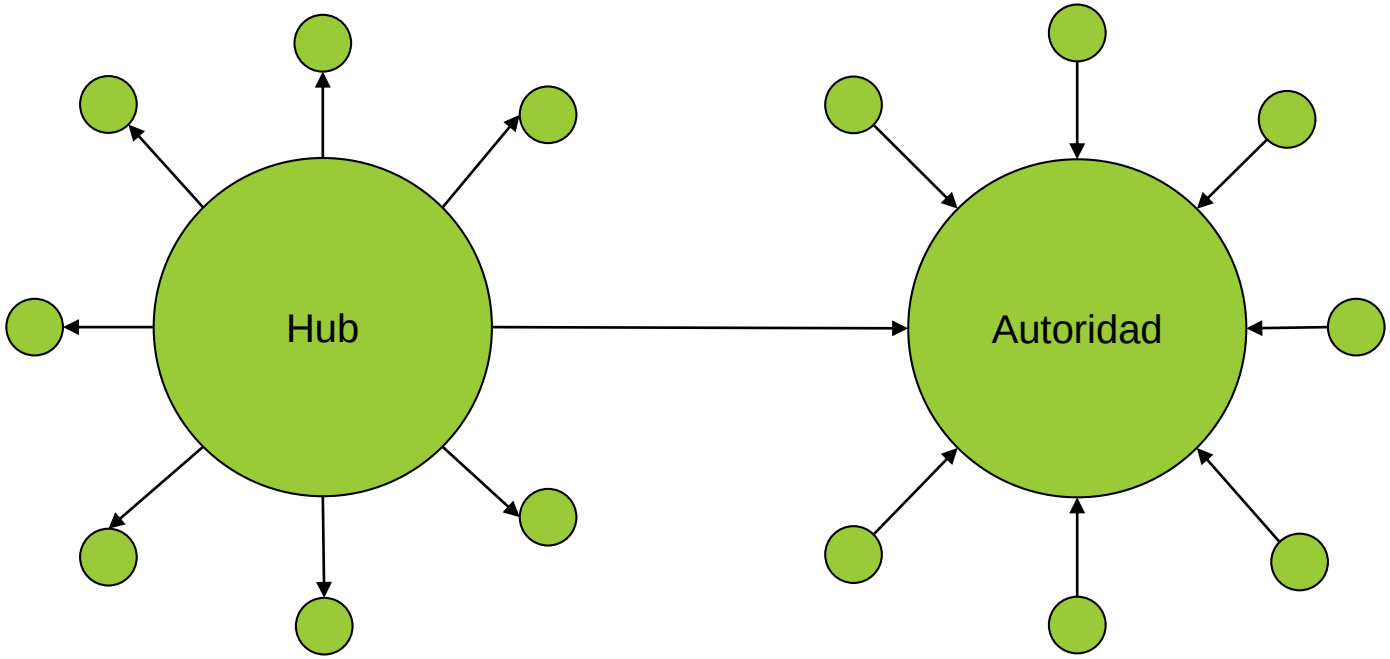
Ranking

HITS

Ranking
HITS

- Algoritmo HITS (Hypertext Induced Topic Search)
- Dependiente de la consulta
  - Tiene en cuenta el conjunto de páginas S que apuntan o son apuntadas por las páginas de la respuesta
- Autoridades: páginas que tienen muchos enlaces apuntándoles en el conjunto S
  - Susceptibles de tener contenido acreditado y por tanto relevante
- Hubs: páginas con muchos enlaces salientes
  - Susceptibles de enlazar contenido relevante similar
- Entre las autoridades y los hubs se cumple una retroalimentación doble:
  - Las mejores páginas vienen acreditadas por los enlaces salientes de buenos hubs
  - Los mejores hubs se forman enlazar buenas autoridades

Ranking



Ranking HITS

- El algoritmo no funciona correctamente con, enlaces inexistentes, repetidos o generados automáticamente
  - Solución: ponderar cada enlace por el contexto que lo rodea
- El conjunto resultado puede incluir páginas no relacionadas con la consulta
  - Solución: asignar una puntuación al contenido de la página (RI tradicional) y combinarlo con el peso del enlace
- Usando esta técnica, la precisión y el recall de los 10 primeros resultados mejorar considerablemente

Ranking

PAGERANK

Ranking PageRank

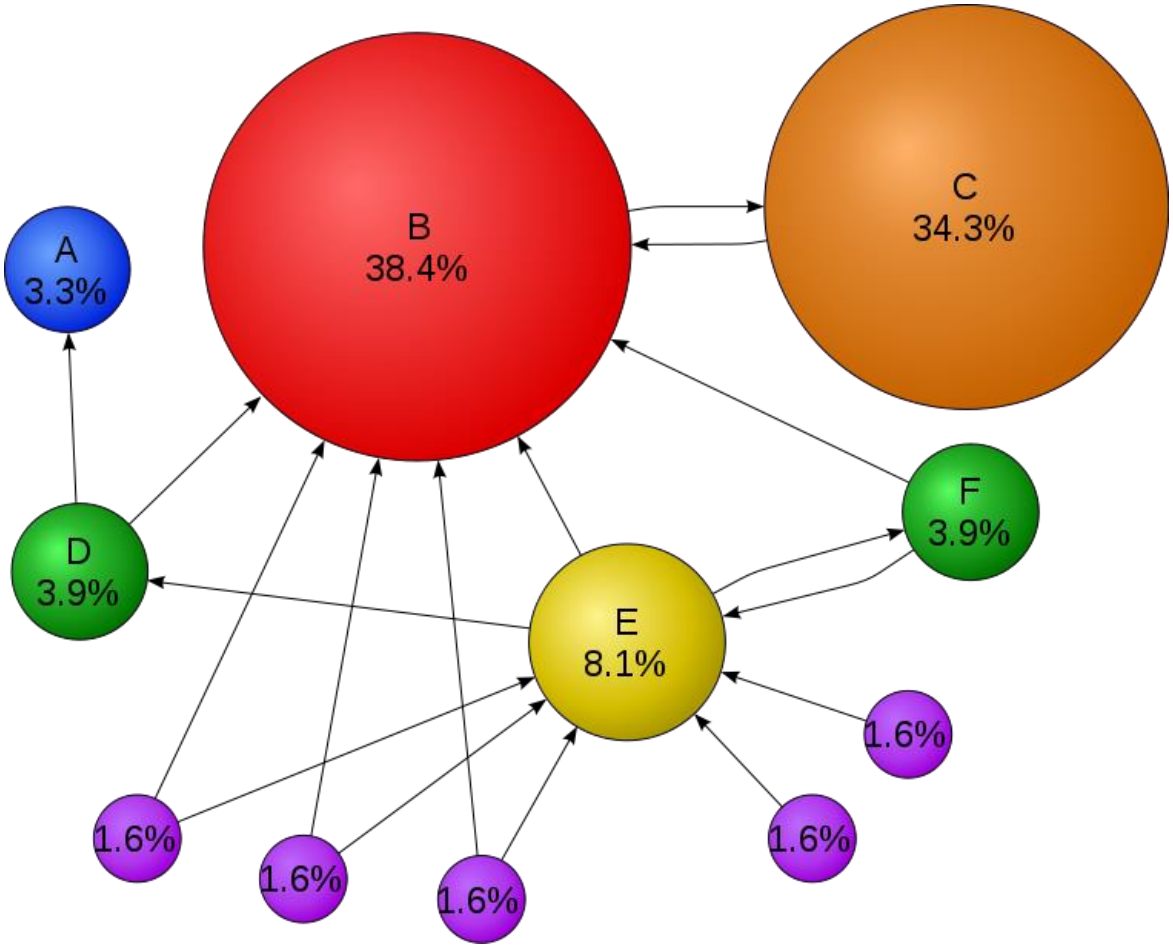
- Algoritmo utilizado por Google
- Su funcionamiento simula a un usuario navegando al azar por la web:
  - Un usuario se encuentra en una página a
  - A continuación, se mueve a una de las páginas enlazadas desde a seleccionando un enlace al azar
  - Después, repite el proceso.
  - Tras un gran número de movimientos, se calcula la probabilidad con la que el usuario visitará cada página
- La probabilidad calculada es una propiedad del grafo
- La web suele contener páginas sin enlaces y páginas que se enlazan a si mismas
  - El usuario puede saltar a otra página con una probabilidad q

Ranking PageRank

$$PR(A) = (1 - d) + \mathrm{d} \sum_{i=1}^{n} \frac{PR(i)}{C(i)}$$

- $PR(A)$ es el PageRank de la página A
- d es un factor de amortiguación (probabilidad con la que se salta a una página al azar en lugar de seguir un enlace)
- $PR(i)$ son los valores de PageRank que tiene cada una de las páginas que enlazan a A
- $C(i)$ es el número total de enlaces salientes de la página i, sean o no hacia A

Ranking
PageRank



Web spam

Web spam

- Muchas personas tienen intereses monetarios en la web
- Los dueños de las webs necesitan que estas aparezcan en las primeras posiciones de los rankings
  - Incentivos económicos
- Web spam: acciones engañosas para escalar posiciones en los rankings
  - Adversarial Information Ranking: área de investigación centrada en el web spam
- Cualquier estrategia de evaluación que cuente características replicables de las páginas webs puede manipularse
- A lo largo de la historia de los motores IR web se han desarrollado numerosas técnicas de web spam
  - *Juego del gato y el ratón*

Web spam

- Técnicas de web spam
  - Página web con un número extremadamente grande de palabras clave
  - Granja de enlaces
  - Estructura compleja de citación entre un conjunto de webs
  - Click spam: robot hacen consultas predefinidas haciendo clic en los enlaces que se quieren promocionar
  - Inyección de código en webs
    - La información mostrada al usuario es diferente de la mostrada a la araña web

Web spam

- No se tiene que confundir web spam con SEO (Search Engine Optimization)
- Las técnicas SEO son legitimas si los webmaster siguen las normas
- Pero, algunas técnicas SEO son directamente web spam
- Si un motor de búsqueda detecta que una web ha sido promocionada mediante web spam, se sacará fuera del índice

Gestión de Datos Web

Gestión de Datos Web

- Asignación de identificadores a los documentos
- Almacenamiento de los metadatos de las web
  - Es necesario una base de datos muy rápida, eficiente y distribuida
    - Google BigTable
    - Hbase: open source
- Compresión del grafo web
  - Las web con contenido similar suelen estar en dominios similares
    - Asignar identificadores similares de modo que parte de ellos sean iguales
- Gestión de datos duplicados
  - Detección de URLs que representan exactamente a la misma página
  - Detección de multiples URLs que enlazan a contenido parcialmente duplicado
  - Reduce el tamaño de la colección