Retorno de información en la web

Las **maquinas de búsqueda en la web** procesan los sitios web y colecciones de documentos

- los documentos son paginas web
- Se mantiene un repositorio indexado de paginas web, usualmente con indices invertidos, que deben ser actualizados regularmente.
- Los resultados a consultas son las pagweb mas relevantes para el usuario.
  Las **máquinas de búsqueda verticales** son personalizadas para tópicos específicos; cubren una colección específica de documentos en la web.

En la web no siempre tenemos acceso a **todos** los documentos disponibles para hacer una búsqueda. Por lo que hay que encontrarlos y construir una colección. Para ello existen los **rastreadores web**.

- Son programas que localizan y recolectan información en la web. Usan los hiperenlaces presentes en documentos conocidos para encontrar otros,
- Se comienza desde un **conjunto semilla** de documentos.
- Los docs recolectados son procesados por un **sistema de indexado** , en donde pueden ser descartados o almacenados como una copia en cache. (toma meses realizar un rastreo)

Como en la web la cantidad de paginas es muy grande, el rastreo se realiza en multiples maquinas en paralelo. De modo que el conjunto de enlaces a ser rastreados se almacena en una bd. Y los nuevos enlaces encontrados en las paginas rastreadas se pueden añadir a este conjunto para ser rastreados a continuación.

Indexado y búsquedas

Como el indexado toma mucho tiempo y se quiere poder procesar muchas consultas al mismo tiempo. Este proceso se ejecuta en multiples maquinas al mismo tiempo.

- Se crea una nueva copia del indice en lugar de modificar el indice viejo.
- El indice viejo se usa para contestar consultas
- Luego de completar la fase de rastreo y antes de volver a comenzar un nuevo indexado, el indice nuevo se convierte en indice viejo.
- Los indices se mantienen en memoria, y las consultas se enrutan a diferentes maquinas.

Relevancia de documentos en la web

Porque usar solo TF-IDF para ordenar respuestas genera problemas? No siempre las paginas que desean ver los usuarios tendrán una frecuencia de términos alta.

- La mayoría se interesa por los sitios populares

- Pueden existir paginas spam, que contienen ciertas palabras sin valor solo para aparecer en las búsquedas.
  Por ello se agrega una nueva forma de medir relevancia de un sitio. Se podría utilizar la **popularidad de un sitio web** como un nuevo parámetro. Pero esto es difícil de medir, por lo que se puede usar el numero de enlaces a un sitio S como medida de popularidad o prestigio del sitio.
- Contar un enlace a S desde cada sitio que lo enlaza a S.

Un problema surge al definir el concepto de sitio, ya que una URL puede contener muchas paginas no relacionadas de variada popularidad.

Por ello se puede combinar TF-IDF con alguna medida de popularidad del sitio para poder obtener una **medición global de la relevancia** de la pagina para una consulta. Luego las de mayor relevancia se devuelven como respuesta top de una consulta.

- **Centro**(hub): es una pagina web o sitio que enlaza una colección de sitios prominentes (autoridades) en un terminado tópico.
  - Una autoridad es una pagina que es apuntada or varios centros buenos
    - Un centro bueno apunta a varias autoridades buenas
- Se busca ordenar por relevancia y autoridad
- Una pagina tiene enlaces **hacia afuera** (otras pag web) y **hacia adentro**(desde otras paginas web)
- No todos los enlaces hacia una pagina son importantes. Un enlace a una pagina desde una fuente creíble es mas importante que un enlace desde una pagina arbitraria.

Page Rank

- es un algoritmo definido por Google como una medida de popularidad basada en la popularidad de las paginas que la enlazan.
- analiza enlaces hacia afuera y adentro
- Se considera a las paginas altamente enlazadas por otras paginas como mas importantes (+ autoridad) qe las paginas con menos.
- pagina P tiene un **rango alto** si la suma de los rangos de las paginas que apuntan a P tiene un valor alto.
- El page rank de una pagina es la probabilidad de que sea visitada en un determinado punto del tiempo.
  - 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.
  - 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 $\delta \in [0, 1]$ y N es el numero de paginas. δ representa la probabilidad de u aso en la caminata aleatoria sea un salto aleatorio.

- El conjunto de ecuaciones se resuelve de manera iterativa,donde se inicia con P[i]=1/N
- Cada paso computa nuevos P[i] con el valor anterior de P.
- La iteracion termina cuando el valor máximo de cambio en un valor de P[i] va por debajo de un cierto valor de corte.

Esto + los valores de TF-IDF de una pagina se combinan para determinar la relevancia de una pagina.
Las maquinas de busqueda llevan la pista de que fraccion de veces los usuarios seleccionan una pagina retornada. Esto tambien aporta a la popularidad de un sitio.

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.

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 computo 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.

A pesar de que las formas en que las empresas integran TF-IDF con alguna medida de popularidad son secretas. Una idea de solución es

- Una formula que combina puntajes es fija y como parámetros toma 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.