Motores de Búsqueda Web
Tarea Tema 1

José Alberto Benítez Andrades

71454586A

Motores de Búsqueda Web
Máster en Lenguajes y Sistemas Informáticos - Tecnologías del Lenguaje en la Web
UNED
15/01/2011

15 de enero de 2011

Tarea 1

Enunciado del ejercicio

La tarea del tema 1 consiste en localizar, entre los proceedings de la conferencia WWW de los últimos tres años (2006-2008), dos artículos relevantes que estudien las estrategias con que los usuarios utilizan los buscadores web o, en general, localizan información en la Web. Debe señalarse, para cada artículo, cuál es su contribución principal y, al menos, dos aspectos potencialmente mejorables del trabajo presentado por sus autores.

1. Resolución

Los proceedings elegidos son los siguientes:

1. Automatic Identificacion of User Intereset For Personalized Search - Fenq Qiu y Junghoo Cho - WWW Conference 2006.

Enlace: http://www2006.org/programme/files/pdf/580.pdf

2. Investigating Behavioral Variability in Web Search - Ryen W. White y Steven M. Drucker - WWW Conference 2007.

Enlace: http://www2007.org/papers/paper535.pdf

1.1.Automatic Identificacion of User Intereset For Personalized Search

En este primer proceeding, sus autores comienzan explicando la necesidad de que los buscadores capten lo que el usuario realmente quiere buscar con un buscador personalizado. Explica con un ejemplo, que dos usuarios pueden buscar con un mismo término cosas diferentes, por ejemplo: alguien que busque en google "office" puede estar buscando una oficina en la que trabajar, sin embargo otra persona que teclea lo mismo, puede estar buscando el programa de Microsoft que tiene ese mismo nombre.

Es una tarea compleja la de poder diferenciar lo anterior, pero ellos dan una serie de pasos a seguir para conseguir el objetivo de crear un buscador personalizado. En primer lugar, se necesita crear un modelo de usuario razonable que capture el historial de cada usuario con sus temas de interés. Basándonos en este modelo, se necesita diseñar un método de aprendizaje que identifique el interés del usuario. Y por último, se necesita desarrollar un mecanismo de ranking que considere el interés aprendido de cada usuario en sus resultados de búsqueda.

El trabajo realizado por Feng Qiu y Junghoo Cho, se basa en la aplicación de un sistema de ranking llamado *Topic-Sensitive PageRank*.

Page Rank y Topic-Sensitive PageRank.

En resumen, el PageRank se basa en un algoritmo de navegación aleatoria por páginas realizando distintas búsquedas, teniendo en cuenta la cantidad de webs que están enlazadas a otras y los enlaces que salen de ellas.

El *Topic-Sensistive PageRank* es una extensión del PageRank que puede dar distintos rankings de las webs para diferentes consultas. Una web A puede tener una puntuación X para una consulta Y y a su vez puede tener una puntuación distinta de X para una consulta Z. Debido a que esta variante de PageRank, contempla puntuaciones según los temas que se buscan, vieron los autores de este artículo interesante basarse en este modelo.

Búsqueda personalizada basada en las preferencias de los usuarios.

En esta parte del artículo, los autores explican cómo utilizan el método de ranking anteriormente explicado para la personalización de las búsquedas. Los autores observan que los usuarios generalmente que tienen preferencias hacia una serie de temas, no les interesan otros temas, con lo cual pueden reducir el conjunto de todas las webs que hay en internet, a un subconjunto más reducido. Por ejemplo, un físico que está interesado en artículos relacionados con las ciencias físicas, por lo general, no va a estar interesado en leer artículos relacionados con los videojuegos.

Representación de la preferencia de los usuarios.

Teniendo en cuenta esto, ellos presentan las preferencias de los usuarios de la siguiente manera:

- Definición 1 (Vector de preferencia de temática) : Un vector de preferencia de temática se define como el conjunto T = [T(1), ... , T(m)] de m-tuplas en los que m es el número de temas a considerar y T(i) representa el grado de interés en el tema i.

Ejemplo 1: Suponemos que hay dos temas : "Ordenadores" y "Noticias", y el usuarios se ha interesado tres veces en "Ordenadores" y una vez en "Noticias". El vector de preferencia es [0.75,0.25].

- Definición 2 (Vector de preferencia de página) : Un vector de preferencia de página se define como un conjunto P = [P(1), ... , P(n)] donde n es el número total de páginas web , P(i) representa el grado de interés de una web i.

En muchas ocasiones este vector puede parecer mejor que el de preferencia temática, ya que, al guardar las páginas web que más visita, se conoce más detalladamente los gustos del usuario. Pero esto no es del todo bueno, ya que, pueden existir en internet muchas webs de la temática que le gusta que sean más interesantes que las que generalmente visita el usuario, por ello es necesario el primer vector.

Modelo de usuario

Para conseguir el vector de preferencia de temática, debemos obtener la información de páginas a las que clicka el usuario, y el grado de interés que tiene el usuario en ellas. Para conseguir esto, primero utilizan el modelo *topic-driven random surfer model.*

- Definición 3 (Topic-driven random surfer model) : Considerando un usuario con vector de preferencia T. Bajo este modelo, el usuario navega por la web en dos pasos. Primero, el usuario elige un tema de interés t para la secuencia de navegación aleatoria con probabilidad T(t). Entonces, con igual probabilidad, el navegador va a una de las webs con temática t. Comenzando por esta página, el usuario realiza una navegación aleatoria, como en cada paso, con probabilidad d, aleatoriamente sigue un enlace de salida en la misma página; con la restante probabilidad 1 - así recoge las que son "aburridas" y vuelve a repetir todos estos pasos con una temática nueva.

Ejemplo : Suponemos que hay 2 temáticas : "Ordenadores" y "Noticias" y un vector de preferencia [0.7,0.3]. Bajo el modelo explicado anteriormente, significa que un 70% de la sesión de navegación aleatoria va a estar dedicada a los ordenadores y un 30% a noticias. Teniendo en cuenta esto, y el método de Ranking elegido (Topic-Sensitive PageRank TSPR), crean un vector de probabilidad de visitas.

- Definición 4 (Vector de probabilidad de visita) : Está definido por un conjunto V = [V(1),.. ,V(n)] en el que n es el número total de páginas y V(i) representa la probabilidad de que el usuario visite esa página i.

Pero para poder conseguir esto, Feng Qiu y su compañero, mejoraron la manera de navegar, dejando de ser aleatorio y pasando a utilizar el modelo *Topic-Driven Searcher Model*.

- Definición 5 (Topic-Driven Searcher Model) : Considerando un usuario con un vector de preferencia T. Bajo un modelo de este tipo, el usuario siempre visita las páginas a través de un buscador en dos pasos. Primero el usuario elige un tema de interés t con probabilidad T(t). Entonces el usuario va al motor de búsqueda y realiza una consulta con la temática elegida t. El buscador devuelve las páginas rankeadas por el TSPR.

Arpendiendo el Vector de preferencia de temática

Basándose en el modelo de usuario anterior y en el método TSPR, se realiza un estudio de las preferencia del usuario. Si por ejemplo, de 10 páginas dadas, el usuario entra 2 veces a una, una vez a otra y ninguna vez al resto, habrá un vector de preferencia del estilo V = [2/3,1/3,0,0...]

Una vez finalizan la explicación de los distintos algoritmos que existen y las personalizaciones que realizan, en el artículo comentan los distintos resultados obtenidos haciendo distintos experimentos y tratando con distintos métodos de PageRank.

Conclusionesde los autores

Después de realizar todas las pruebas, comentan sus autores como conclusión final, que el sistema inicial de PageRank se puede mejorar notablemente realizando unos cambios como los realizados por ellos mismos para realizar este experimento. Señalan que google ha comenzado a desarrollar un servicio de búsqueda personalizada que parece estimar el interés del usuario por distintas consultas anteriores.

Mejoras posibles de este estudio

El estudio realizado por Feng Qiu y Junghoo Cho me ha parecido bastante interesante y muy bien planteado. Quizá yo hubiera planteado el tema de las búsquedas personalizadas añadiendo una serie de parámetros más para mejorar los resultados.

Una posibilidad, es la adición de un sistema de votos de web por consulta, de manera que, el propio usuario, de una lista de webs dadas realizando una búsqueda, pueda votar del 0 al 10, el contenido de la página web que ha consultado, indicando si la información que contiene le ha servido o no.

Tengo entendido que en la actualidad, Google está integrando algo parecido a este sistema sobre todo para las páginas web de distintos negocios, teniendo en cuenta los votos positivos y los negativos. También leí, que hasta hace poco, en este sistema que estaba en pruebas, tenían un fallo bastante importante, y era, que sólo tenían en cuenta el número de votos y comentarios que tuviera una web, sin determinar si los votos eran positivos o negativos.

1.2.Investigating Behavioral Variability in Web Search

En este artículo, Ryen W. White y Steven M. Drucker tratan el tema de las búsquedas personalizadas desde un estudio longitudinal basado en logs que investiga la variabilidad en la interacción de las personas. Analizan la interacción de más de dos mil usuarios a lo largo de 5 meses con el objetivo de caracterizar diferencias en sus interacciones dentro de y entre los usuarios y dentro de y entre las consultas que ellos realizan.

Introducen el tema indicando que es muy importante en la época en la que estamos, la búsqueda de información desde los distintos buscadores que existen en la actualidad. Indican que hay muchas maneras de buscar las cosas y que cada usuario realiza las búsquedas de una manera particular, por ello quisieron realizar el estudio con un grupo numeroso de personas.

Métodos que utilizaron para su estudio.

Para realizar el estudio basado en logs, hacían una pregunta a los usuarios que instalaron la aplicación en la cual pedían el consentimiento de guardar logs con las interacciones que hicieran los usuarios en la web: 3291 personas aceptaron. El periodo en el que capturaron las actividades de los usuarios fue desde Diciembre de 2005 hasta Abril de 2006.

El mecanismo utilizado se instalaba como un plugin del navegador Microsoft Internet Explorer que escribía una entrada en un servidor remoto cada vez que se abría una web en el navegador. Se almacenaba un identificador por cada usuario (anónimo), una marca de tiempo por cada página vista y la url de la página visitado. Por razones de privacidad, no almacenaban el contenido de las páginas vistas sobre conexiones seguras.

En los 5 meses de estudio, los 3291 usuarios visitaron millones de páginas y estuvieron muchas horas en internet.

Camino d extracción de búsqueda

Los logs de cada participante se iban agrupando dependiendo de la información de navegación. Dentro de cada instancia de navegador, los participantes eran representados en forma de una ruta constante que se referencian a un *browser trail*, desde la primera a la última página web visitada en el navegador. Dentro de algunos de estos caminos de búsqueda se producían por motores de búsqueda como Google, Yahoo, MSN Search y Ask.

Los caminos de búsqueda originados con una búsqueda directa y proceden hasta un punto de terminación donde es asumido que el usuario ha completado su información de actividad de búsqueda. Los caminos pueden obtener múltiples interacciones de búsuqeda y deben contener páginas que son o: resultados de búsqueda, visitas a páginas de inicio de motores de búsqueda, o conectadas con una página de resultados de búsqueda vía hipervínculos. Para reducir la cantidad de ruido de las páginas no relacionadas a las tareas de búsqueda activas introdujeron algunas actividades de terminación que usaban para determinar los puntos finales de los caminos de búsqueda:

- Volver a la página de inicio: Volviendo a la página de inicio se asumía que era el fin de un camino de búsqueda. Aunque no tenían acceso a la configuración de los navegadores de los usuarios, analizaban sobre la página que aparecía primeramente cuando se abría un navegador.

- Comprobar el email o autentificarse en un servicio: Comprobaban los e-mails basados en web o las autentificaciones a servicios web como myspace.com o del.ico.us para determinar que finalizaba el camino de búsqueda.
- Escribir la URL o visitar páginas que están en marcadores: Introducir una URL de forma directa en la barra de navegación o seleccionar un marcador, determinaba el fin de un camino.
- Timeout de una página: Si el tiempo de muestra de una página excedía los 30 segundos, era el fin de un camino.

Para representar estos caminos, utilizaban grafos, en los cuales se veían los caminos a distintas webs y las posibles alternativas para llegar hacia ellas.

Selección de consultas.

Durante el estudio, los participantes realizaron alrededor de tres millones de consultas y siguieron aproximadamente medio millón de caminos de búsqueda donde por lo menos un resultado fue clickado. El 15.6% de las consultas aparecieron solo una vez en las consultas de los participantes. Aunque la presencia de una distribución Zipfian es común en los logs de búsqueda web, fue potencialmente problemático en el análisis. De cada conjunto de consultas, seleccionaron un subconjunto de consultas que habían sido realizadas al menos 15 veces y por 15 participantes. Esto daba suficientes datos y solapaba entre participantes y consultas. El conjunto resultante de las consultas representa el 10.2% de las consultas iniciales y contiene las 385 consultas más populares. Idealmente podían haber utilizado todas las consultas en el análisis que realizaron. Sin embargo, con la cantidad de consultas que tenían almacenadas, era algo inviable.

Findings

En los 5 meses del estudio, los usuarios vieron 80 millones de páginas web, 12.5% seguían caminos de búsqueda que ya habían sido hechos anteriormente. Categorizaron las webs en 2 grupos : search (S) y browse (B). Entre estos tipos de páginas había dos tipos de transiciones: forward(f) y backward(b). Así que había finalmente 4 categorías : forward-to-search, backward-to-search, forward-to-browse, backward-to-browse.

Menos de un tercio de interacciones es con páginas de resultados y el resto es con páginas que siguen un camino de hipervínculos de una página de resultados. Estos resultados no estresaban la importancia de las interacciones post-consulta, pero también demostraban el volumen de interacción que había para el análisis cuando ambos, navegación y búsqueda eran considerados.

Ellos enfocan el estudio en la variabilidad de la interacción para consultas y usuarios. Para esto y todos los análisis de subsecuencias ellos agrupaban en dos caminos los datos: por el usuario y por la consulta.

Variabilidad de usuario

Dividen el análisis en tres fases : estudian las diferencias en los patrones de interacción, entonces la inclusión de características adicionales del camino y finalmente la variación en los dominios participantes visitados.

15 de enero de 2011

1. Diferencias en patrones de interacción.

Para caracterizar los caminos de búsqueda necesitaban un camino para representar las consultas de los usuarios que permitan comprara las interacciones. Modificaron un método usado en trabajos parecidos que representaba los caminos como cadenas y entonces computaban la distancia de Levenshtein entre caminos que representaban el mismo camino (LD). LD es un método para juzgar la proximidad de dos cadenas de longitud arbitraria.

Ejemplo: Dado los 3 caminos siguientes, usaremos el enfoque descrito anteriormente para determinar los caminos más representativos:

Paso 1: Representar caminos como cadenas.

1: S1->S2->S3->S2->S5->S6 = SSBbSBS

2: S1'->S2'->S3'->S2'->S5'->S1'->S6 = SBBbBSbSS

3: S1''->S2''->S3''->S4''->S5'' = SBBBB

Paso 2: Calcular la distancia media entre cadenas:

Desde el camino 1	Desde el camino 2	Desde el camino 3
LD(1,2) = 4	LD(2,1) = 4	LD(3,1) = 4
LD(1,3) = 4	LD(2,3) = 4	LD(3,2) = 4
Average = 4	Average = 4.5	Average = 4.5

Paso 3: Seleccionar el camino más representativo.
El camino 1 tiene la distancia mínima más baja = 4.

Las clases de usuario que han realizado este estudio, son 2 : navegantes y exploradores.

Navegantes (varianza baja): Estos usuarios tienen patrones de intreacción consistentes en los caminos que ellos siguen. Que es, muchos de sus caminos de búsqueda son similares cuando ellos son reducidos para la representación usada para computar LD. Los usuarios cuyas interacciones son consistentes semejantemente interactuadas en un camino particular. Los caminos de este tipo de usuarios presentan 3 atributos: ellos realizan muchas desviaciones o regresiones, aparecen para hacer frente a problemas secuenciales y ellos revisitan dominios.

Exploradores (varianza alta): Estos usuarios tienen unos patrones de interacción variables en los caminos que ellos siguen. Esto es que muchos de sus caminos de búsqueda para cada uno de los usuarios es diferente cuando ellos reducen la representación al computar LD. Estos usuarios tienen tres atributos: tienden a la rama con frecuencia, realizan muchas consultas por cada sesión y visitan muchos dominios nuevos.

Diferencias en las características de los caminos

Para que los análisis de los caminos fueran más completos, tuvieron en cuenta una serie de características:

- Tiempo: Cantidad de tiempo que se gasta en cada camino.
- Número de consultas: número de consultas que se han realizado en cada camino.
- Número de pasos: número de páginas vistas en cada camino incluyendo todas las búsquedas y revisitas.
- Número de revisitas: número de revisitas a páginas vistas en caminos anteriores.
- Número de ramas: número de veces que un tema revisita una página anterior en un camino y procede a ver otra página.
- Media de longitud de una rama: media de número de pasos en cada rama en el camino.

Teniendo en cuenta estos factores, se llego a los siguientes datos respecto a los factores que más provocaban una varianza entre los usuarios:

- *Forward and backward motion:* el factor con más poder, contribuyó al 52.5% de la varianza. Aparece para representar una dimensión básica que relata el clickthrough y las operaciones *back*.
- *Branchiness:* este factor cuenta con el 17.4% de la varianza y representa la medida para que los usuarios sigan subramas dentro de una rama de búsqueda.
- *Tiempo*: Este cuenta con el 10.7% de la varianza y representa la cantidad de tiempo para atravesar un camino de búsqueda.

Diferencias en los dominios visitas

Otro dato importante que tuvieron en cuenta para el estudio, es que, los que visitan muchas veces un mismo dominio, suelen ser usuarios que no muestran mucha varianza en sus patrones de interacción.

Conclusiones finales de los autores

Los autores de este artículo describieron un estudio de investigación de gran escala basado en logs sobre la variabilidad en las actividades de búsqueda de los usuarios. Sobre dos mil participantes tomaron parte de este estudio de 5 meses. Las búsquedas sugerían que hay usuarios y consultas cuya interacción es particularmente consistente y particularmente variable, y que, esto es posible para caracterizar muchas características de variación con un pequeño número de dimensiones que pueden ser útiles en el diseño de la interface. Además para soportar las masas, quienes aparecen para interactuar, los motores de búsqueda web deben proveer herramientas que están basadas en las necesidades de los usuarios que exhiben búsquedas regularmente. En un futuro trabajo, sus autores pretenden crear un catálogo de patrones de uso web que predigan lo que el usuario quiere buscar.

Mejoras posibles de este estudio

Este estudio realizado por Ryen W. White y Steven M. Drucker ha sido muy interesante, aunque me pareció más interesante el primer proceeding que comenté en este trabajo. Su manera de recoger la información fue bastante buena, pero iba centrado a un público bastante exclusivo, ya que, el plugin que realizaron era únicamente para Microsoft Internet Explorer. Creo, desde mi punto de vista, que el público que utiliza Microsoft Internet Explorer suele ser un público que no está muy avanzado en cuanto a lo que Internet se refiere, ya que, suele ser gente que desconoce de la existencia de otros navegadores alternativos, y esto conlleva a que tienen un conocimiento muy reducido de todo lo que podemos hacer con los ordenadores.

Así, el estudio de las páginas que revisita cada persona, o por los que pasan hasta llegar a un resultado, seguramente sean distintos a los que un usuario avanzado realice. Por ello pienso que podían haber integrado otros plugins en los navegadores alternativos que existían en su momento.

Además, pienso que el algoritmo elegido para controlar los caminos de búsqueda que realizaba cada usuario, puede ser bastante mejorable para recoger unos datos quizá más interesantes de los recogidos en el estudio.