Minería de la Web
Tema 1

José Alberto Benítez Andrades

71454586A

Minería de la Web
Máster en Lenguajes y Sistemas Informáticos - Tecnologías del Lenguaje en la Web
UNED
07/12/2010

7 de diciembre de 2010

Tema 1

1. Problemas que surgen al interactuar con la web.

- Existen diferentes tipos de problemas a la hora de interactuar con la web:

- Usabilidad: Una web debe ser lo suficientemente intuitiva como para que cualquier tipo de usuario acceda a cualquier parte de la web sin dificultades. Esto quiere decir, que tenga una navegabilidad clara y sencilla, con un contenido coherente, títulos concretos y correctos, y descripciones correctas. De lo contrario, cuando un usuario entre a nuestro sitio web, no va a saber qué queremos.
- Accesibilidad: Debemos hacer que el código de nuestra web cumpla con los estándares de calidad, para que, de esa manera, cualquier persona con cualquier tipo de discapacidad, pueda disponer de la información que se encuentra en la web sin ningún problema.
- Búsquedas en los buscadores actuales: Actualmente es complicada la tarea de buscar información en la red, ya que, los buscadores no comprenden semánticamente los términos que buscamos, con lo cual, encontramos en muchas ocasiones páginas que no contienen la información que realmente queremos buscar.

- Los diseñadores y programadores de páginas web, deben contar con una serie de factores importantes a la hora de crear una página web. Por ejemplo, las webs gubernamentales, están obligadas por ley a tener un nivel de accesibilidad Doble AA, medido mediante unos validadores proporcionados a lo largo de la red, como por ejemplo:

- http://validator.w3.org , desde este validador, podemos ver si nuestra página cumple con los estándares de HTML o tiene algún tipo de error.

- http://www.cynthiasays.com/ , desde esta web, medimos el nivel de accesibilidad web, proporcionándonos información muy importante para controlar si nuestra web es totalmente accesible o no.

2. Breve definición de Minería de la web y de

- La minería de la web, engloba todos los elementos que definimos en el punto actual, el crawling, la búsqueda en la web, la minería del contenido en la web, la minería de uso en law eb y la minería de estructura de la web.

I. Crawling.

- El término de 'crawling' se utiliza para referirnos a la técnica de rastreo web por parte de los buscadores. Es decir, la acción que realizan los buscadores de entrar a todas las webs para obtener toda la información que contienen para un posterior análisis, es lo que llamamos **crawling**.

II. Búsqueda en la web.

- Búsqueda en la web, se define como, el proceso de buscar en Internet la información que necesitamos encontrar. Por ejemplo, si queremos conocer cuál es la capital de Lituania, en el buscador debemos teclear "capital de Lituania" , ya que, el buscador no entiende la semántica de la frase, simplemente entiende de términos clave.

III. Minería de contenido de la web (minería de texto).

- Proceso mediante el cual se recuperan todos los datos de todas las webs que hay en internet, se analizan y se crean estadísticas sobre ellos.

IV. Minería de uso de la web.

- Proceso que se encarga de recopilar toda la información referente a las sesiones cliente-servidor, es decir, los logs de los servidores, las cookies y otros tipos de datos que se comparten entre el servidor y el cliente que ve la web.

V. Minería de estructura de la web.

- Proceso mediante el cual, utilizando la teoría de grafos, se examinan los nodos que conectan unos hipervínculos con otros, a lo largo de toda la red de Internet.

VI. Dinámica de la web.

- La dinámica de la web es el término que se utiliza para englobar las distintas técnicas que se utilizan para crear una página web, la mejora de las estructuras, modelos, etcétera.

3. Qué problemas abordan cada una de las áreas anteriores.

I. Crawling.

- En el 'crawling' existen distintos problemas que por lo general para todas las áreas suelen ser parecidos:

- Uso del ancho de banda: cuando el rastreador se pasea por todas las páginas, hay un consumo alto de ancho de banda, tanto por parte del rastreador como por parte de la web rastreada.
- Control de acceso al servidor: debe existir un fichero llamado **robots.txt**, en el cual se deje claro a qué directorios y qué ficheros deben ser rastreados y cuáles no, ya que, en muchas ocasiones, rastrear documentos y carpetas con contenido no relevante, retrasa la tarea de los rastreadores, incrementa el consumo de recursos por parte del servidor, y en definitiva, es negativo en todos los sentidos.
- Control de acceso a los recursos: mediante meta-etiquetas, debemos dejar bien claro en cada página, si queremos que sigan los enlaces que hay en la web o no.
- Para que una web pueda ser rastreada, debe contar los puertos necesarios abiertos.
- Otro problema es el cuello de botella que se forma por culpa de los servidores DNS y por los posibles fallos temporales que pueden tener.
- Las redirecciones a páginas de error, deben estar bien implementadas, ya que sino, el rastreador puede entrar en un bucle peligroso.
- La codificación HTML con errores, tags vacíos, mezclar las comillas, no marcar las etiquetas de título correctamente, y otro tipo de errores en la codificación, influyen negativamente en los rastreos.
- El tamaño de las URLs y el formato, influyen a la hora del rastreo de páginas.
- Los contenidos duplicados son también un serio problema
- Los distintos algoritmos de rastreo que existen y que se han ido mejorando a lo largo de la evolución del mundo web, como por ejemplo el algoritmo de *Information Retrieval*, los distintos tipos de aprendizaje automático, etcétera.

II. Búsqueda en la web.

- En la búsqueda en la web, los problemas que existen son los siguientes:

- Mala interpretación de los términos clave: El usuario en muchas ocasiones, no sabe cómo buscar la información que necesita, ya que, piensa que el buscador sabe interpretar de forma semántica lo que quiere. Por ejemplo, si queremos buscar cómo hacer una pizza que esté buena, si el usuario introduce una búsqueda del estilo " qué elementos necesito para hacer que una pizza esté realmente buena" lo más probable es que no encuentre lo que quiere, deberá saber elegir los términos que tiene escribir para encontrar lo que busca.
- Enlaces a páginas que no contienen la información correcta: Todavía hoy en día, ocurre en muchas ocasiones, que hay webs posicionadas en los primeros lugares, que no tienen información correcta respecto al tema que estamos buscando.

III. Minería de contenido de la web (minería de texto).

- Proceso mediante el cual se recuperan todos los datos de todas las webs que hay en internet, se analizan y se crean estadísticas sobre ellos.

IV. Minería de uso de la web.

- Los problemas que puede encontrar principalmente en la minería de uso de la web son los siguientes:Proceso que se encarga de recopilar toda la información referente a las sesiones cliente-servidor, es decir, los logs de los servidores, las cookies y otros tipos de datos que se comparten entre el servidor y el cliente que ve la web.

- Problemas de red: Pueden existir problemas por parte del servidor donde se encuentra alojada la web que está rastreando, como por ejemplo, una caída en la red del hosting, un fallo en las DNS temporal o cualquier otro error de hardware.
- Problemas de código enviado al usuario por parte del servidor web: Si tenemos un error 404, de objeto no encontrado, pero la web está mal programada y en lugar de enviar un error 404, envía un código de recepción 200 OK, o por ejemplo, si tenemos que realizar un redireccionamiento de tipo 301, pero lo indicamos como un 404.

V. Minería de estructura de la web.

- A la hora de realizar la minería de estructura de la web, podemos encontrar problemas como los siguientes:

- Webs no indexadas: Si creas una web, y no indicas a ningún buscador de su existencia, tendrá problemas para realizar la minería de estructura de la web, ya que, no encuentra la web en internet y no entraría nunca.
- Desaparición de webs por cierre de las mismas: Si unas webs que existían, se cierran, a la hora de realizar la minería de estructura, surge un problema que conlleva a fallos de rastreo y pérdida de tiempo intentando analizar una web que ya no existe.

VI. Dinámica de la web.

- El dinamismo de la web, la evolución de los lenguajes de programación web, el cambio de los distintos estándares a lo largo del paso de los años, son un problema bastante grave para todos los rastreadores, ya que, tienen que ir mejorando continuamente el algoritmo de inserción, análisis y modificación de enlaces en los directorios y almacenes de datos.

4. Qué otras áreas de investigación están relacionadas con Minería de la web.

- Otras áreas de investigación que estén relacionadas con la minería de la web, pueden ser por ejemplo:

- El almacenamiento de datos.
- Técnicas de posicionamiento SEO.
- Inteligencia Empresarial.
- Gerencia de la información y análisis de datos.
- Procesamiento de base del conocimiento conceptual.

5. Qué conferencias internacionales tienen relación con Web Mining.

- Después de realizar unas búsqueda por internet he podido encontrar las siguientes conferencias internacionales:

- WSDM (Web Search and Data Mining) Conference - http://www.wsdm-conference.org/2010/
- El comité organizativo está formado por integrantes de Microsoft, Google, Yahoo y Cornell.

- International Conference on based Web-Learning - http://iic.shu.edu.cn/icwl2010/ - La última edición se está celebrando estos días, del 7 al 11 de Diciembre en Shanghai.

- International Workshop on Cognitive-based Interactive Computing and Web Wisdom http://iic.shu.edu.cn/icwl2010/workshop/cicw2010/ - También se está celebrando actualmente la última edición en Shanghai del 8 al 10 de Diciembre.

- IEEE International Conference on Data Mining - http://datamining.it.uts.edu.au/icdm10/ - La próxima edición es del 14 al 17 de Diciembre de este año 2010, en Sydney, Australia.

- WKDD (Conference on Knowledge Discovery and Data Mining) - http://www.apesrc.org/wkdd2011/ - La siguiente cita van a ser los días 10 y 11 de Abril de 2011 en las Maldivas.

- SIAM International Conference On Data Mining - http://www.siam.org/meetings/sdm11/ - La siguiente edición es en los días 28 al 30 de Abril de 2011 en Hilton Phoenix East / Mesa, Arizona USA.

- ADMA Advanced Data Mining and Applications - http://arnetminer.org/html/adma/html/ - La siguiente edición se celebra durante los día 16 a 18 de Diciembre de 2011 en Tsinghua University, Beijing.

6. Lista de referencias utilizadas.

- Kosala, R. and Blockeel, H. Web Mining Research: A Survey. ACM SIGKDD Explorations, Newsletter of the Special Interest Group on Knowledge Discovery and Data Mining. 2000.
- Chakrabarti, S. Data Mining for hypertext: a tutorial survey. ACM SIGKDD Explorations, Newsletter of the Special Interest Group on Knowledge Discovery and Data Mining, 2000.
- Ricardo baeza Yates. Excavando la web. El profesional de la información. v13, n1, 2004.
- R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In Proc. of the 20th VLDB Con- ference, pages 487{499, Santiago, Chile, 1994.
- S. Agrawal, R. Agrawal, P.M. Deshpande, A. Gupta, J. Naughton, R. Ramakrishna, and S. Sarawagi. On the computation of multidimensional aggregates. In Proc. of the 22nd VLDB Conference, pages 506{521, Mumbai, India, 1996.
- R. Armstrong, D. Freitag, T. Joachims, and T. Mitchell. Webwatcher: A learning apprentice for the world wide web. In Proc. AAAI Spring Symposium on Information Gathering from Heterogeneous, Distributed Environments. 1995.
- M. Balabanovic, Yoav Shoham, and Y. Yun. An adaptive agent for automated web browsing. Journal of Visual Communication and Image Representation, 6(4), 1995.
- A. Z. Broder, S. C. Glassman, M. S. Manasse, and G Zweig. Syntactic clustering of the web. In Proc. of 6th International World Wide Web Conference, 1997.

7. Lista de enlaces utilizados.

Los enlaces utilizados para poder contestar a las preguntas referentes a la minería web y derivados, han sido los siguientes:

http://www.daedalus.es/fileadmin/daedalus/doc/I%2BD/DAEDALUS-RP-IBERAMIA_2002_Mineria_Web.pdf

http://serviciosdelaweb.uimp20.es/

http://ants.dif.um.es/staff/juanbot/ml/files/20022003/webmining.pdf

http://elprofesionaldelainformacion.metapress.com/app/home/contribution.asp?referrer=parent&backto=issue,2,10;journal,42,74;linkingpublicationresults,1:105302,1

http://www.lsi.us.es/redmidas/CEDI/papers/189.pdf

http://www.iadis.net/dl/final_uploads/200607L024.pdf

Para poder encontrar las distintas conferencias que hay internacionales, además de Google, he utilizado la siguiente web:

http://www.allconferences.com/