Arquitectura de un Gestor de Noticias

Fernando M. Sagui Ana G. Maguitman Guillermo R. Simari

Laboratorio de Investigación y Desarrollo en Inteligencia Artificial (LIDIA)[1]
Departamento de Ciencias e Ingeniería de la Computación
Universidad Nacional del Sur
Av. Alem 1253, (B8000CPB) Bahía Blanca, Argentina
Tel: (0291) 459-5135 / Fax: (0291) 459-5136
e-mail: {fms, agm, grs}@cs.uns.edu.ar

Resumen

En la actualidad, las formas más comunes de acceder a las noticias en la Web son: (1) a través de la búsqueda por palabras claves, (2) mediante la navegación de directorios, o (3) visitando páginas dinámicamente generadas basadas en el perfil del usuario. Ninguna de estas formas de acceso permite realizar un manejo cualitativo del contenido de las noticias. Por ejemplo, los mecanismos más usuales para la búsqueda y presentación de noticias no incorporan nociones de validez o confiabilidad. En este trabajo describimos la arquitectura de un gestor de noticias que facilitará el manejo cualitativo de la información. Nuestra meta es agregar un nivel adicional a los gestores básicos, donde se aplicarán nuevas heurísticas, se integrarán noticias basadas en fuentes antagonistas y donde será posible fundamentar posiciones contrapuestas. Esto nos permitirá integrar un sistema de recomendación de noticias con sistemas dialécticos, lo cual facilitará un seguimiento más sostenido y amplio de las noticias de interés para el usuario.

1. Introducción

De acuerdo a varios estudios e informes realizados por Nielsen / NetRatings [5], la búsqueda y lectura de noticias en diarios y otros medios de información se ha vuelto una de las actividades más importantes dentro la Web. Tanto la cantidad de usuarios como el tráfico en los principales buscadores de noticias (Yahoo! News, Google News) se ha visto incrementado. La gran cantidad de noticias online refleja la necesidad que tienen los usuarios de estar informados. Además, el éxito que han tenido los buscadores de noticias durante los últimos años marca la necesidad de obtener información y opiniones pluralistas.

En la actualidad, existe una gran cantidad de buscadores de noticias comerciales que están disponibles desde hace unos años. Entre los más utilizados se encuentran Google News [2], Yahoo! News [3], MSNBC [1], etc. Aunque ninguno de ellos da a conocer mediante publicaciones la manera en que ordenan las noticias, es evidente que toman en cuenta factores tales como la novedad, la fuente a la que pertenece el artículo y la cantidad de veces que se repite en diferentes medios.

Una revisión de la literatura actual nos permite identificar una serie de desarrollos en el área de buscadores de noticias. NewsInEssence [14] es un sistema que busca y agrupa en clusters noticias relacionadas. QCS [9] es una herramienta que hace más eficiente la tarea de agrupar y categorizar los

[1]Las investigaciones realizadas en el LIDIA son financiadas por la Universidad Nacional del Sur, por la Agencia Nacional de Promoción Científica y Tecnológica (PICT 2002, Nro. 13096) y por el Consejo Nacional de Investigaciones Científicas y Técnicas (PIP 5050).

documentos en tópicos. En [12], los autores muestran cómo buscar artículos en la web mientras se transmiten noticias en televisión. En [15] se propone una herramienta para extraer noticias de sitios web de manera automática. En [10] se propone y analiza NewsJunkie, un sistema que personaliza las noticias identificando las que son primicia en el contexto del usuario. Velthune [11] es un motor de búsqueda de noticias que extrae información tanto de la Web como de News feeds.

Al momento de definir los servicios que proveerá un gestor de noticias debemos tener en cuenta ciertas características comunes a la mayoría de los lectores:

- **Primicias.** Los lectores quieren recibir información sobre los eventos tan pronto como los mismos ocurran.
- **Relevancia.** Los lectores no quieren ser distraídos con información inútil. Sólo desean información que les resulte de interés.
- **Pluralismo.** A los lectores puede interesarles recibir información de diferentes fuentes sobre una misma noticia, especialmente si dicha información resulta conflictiva. También puede interesarles obtener información extra que fundamente ciertas posiciones.

Las fuentes de noticias han evolucionado a lo largo de los años. Actualmente, no sólo utilizan páginas web para brindar información, sino que también publican utilizando gran cantidad de meta-datos en sus sitios. Esto nos brinda una fuente de recursos importante y que podemos aprovechar para desarrollar herramientas que utilicen esta información que, a diferencia de la web clásica (html), nos brinda recursos en formatos estándares y correctamente clasificados. Los meta-datos más comunes que permiten realizar un manejo más cualitativo de la información son: prioridad, palabras claves, fecha, autor, agencia, categoría (internacional, deportes, economía, último momento, etc.). Además, de estos meta-datos típicos es factible contar con anotaciones provistas por los lectores mediante los procesos colaborativos que facilita la Web 2.0.

Como primer etapa de nuestra investigación estamos desarrollando un Gestor de Noticias (News Manager) que contará con la capacidad para la descarga, clasificación e indexado de noticias. El paso siguiente en nuestra investigación será el agregado sucesivo de niveles adicionales destinados a proveer funcionalidades tales como:

- Sensibilidad a la tarea del usuario para seleccionar noticias basadas en diferentes contextos temáticos.
- Manejo de preferencias para lograr un modelo que se adapte al perfil del individuo que interactúa con el sistema.
- Incorporación de nociones de calidad y validez, así como otros criterios que permitirán realizar un análisis cualitativo del contenido de las noticias.
- Agregado de mecanismos que faciliten las anotaciones semánticas e integración con ontologías.
- Interacción con componentes inteligentes que cuenten con capacidades multilingües y multiculturales.

Cabe mencionarse que pondremos especial atención en el diseño de algoritmos y criterios de ranking. Los métodos más difundidos para generar un ranking de las páginas web (PageRank, Hits, etc.) no siempre serán válidos en nuestro modelo. Las primicias tendrán un peso mayor que las noticias antiguas, pero es probable que las mismas no se encuentren enlazada y que no haya páginas que las apunten.

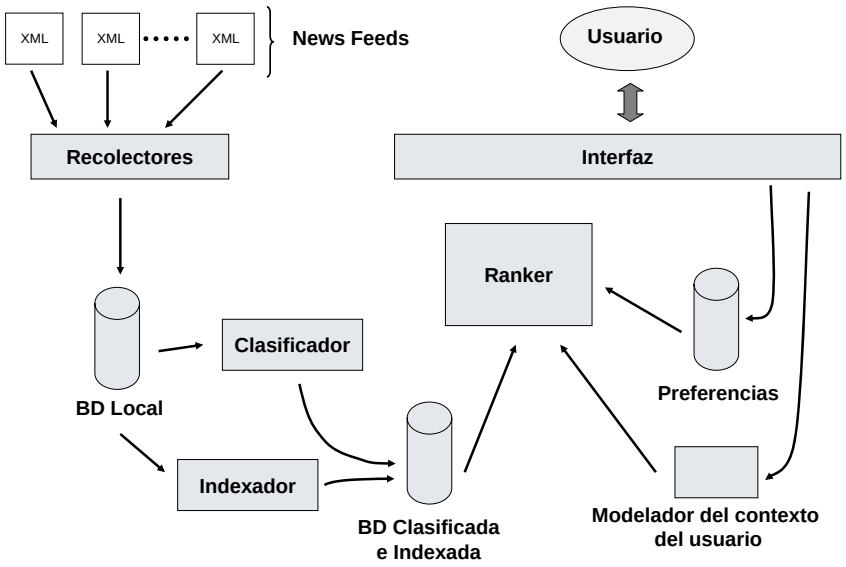

Figura 1: Arquitectura del Gestor de Notiicas

2. Arquitectura

A continuación describimos la arquitectura de nuestro Gestor de Noticias, la cual se muestra esquemáticamente en la Figura 1. El sistema accede a repositorios de noticias, genera bases de datos locales y utiliza diversos criterios para procesar estas noticias y presentarlas a los lectores. La fuentes de datos utilizadas se describen a continuación:

- **Web Feeds.** Las noticias son extraídas de diversas fuentes de información. Utilizaremos las fuentes que publican el contenido de su sitio en alguno de los formatos de "syndication" más usados: `Atom` [4] o `RSS` [6]. En general, la mayoría de los medios utiliza los web feeds para publicar un resumen de los cambios más recientes que se han producido en el contenido del sitio. Actualmente este mecanismo es usado por la mayoría de los diarios para publicar sus headlines rápidamente. El Content syndication ha crecido en gran escala desde el año 2001 como medio de publicación de información de diversa índole. Los feeds proveen meta-información sobre el sitio que se está visitando y son herramientas muy útiles para el manejo de información de forma estructurada. Tanto `RSS` como `Atom` se han impuesto dentro de la Web como estándares utilizados por casi todos los diarios y fuentes de información del mundo. Ambos formatos están codificados en `XML`, razón por la cual pueden ser leídos utilizando cualquier plataforma, browser o aplicación.
- **Base de Datos de Noticias.** Esta base de datos contiene las noticias almacenadas localmente. La misma es periódicamente actualizada por los recolectores (descriptos más abajo), así como también indexada y clasificada para permitir un acceso temático a la información disponible localmente.
- **Base de Datos de Preferencias de usuarios.** El sistema mantiene la historia de las preferencias de los usuarios con el fin de optimizar y ajustar los resultados de acuerdo a sus prioridades. En un principio, las preferencias deberán ser explícitas. Eventualmente esperamos desarrollar un esquema proactivo que genere un perfil del usuario mientras el mismo interactúa con el sistema. Contar con una base de datos con preferencias de los usuarios nos permitirá implementar un sistema de recomendación que utilice nociones cualitativas basado en el esquema propuesto en [8].

Los módulos encargados de facilitar el acceso, procesamiento y presentación de noticias son los siguientes:

- **Recolectores.** Son los encargados de descargar las noticias de un conjunto predefinido de fuentes. Su trabajo consiste en mantener la base de datos local actualizada, accediendo a las fuentes de datos, observando cambios y descargando la información más reciente.
- **Indexador.** El módulo indexador será el encargado de identificar términos importantes que describan las noticias. El mismo tomará aquellas noticias descargadas por los recolectores y generará un índice que estará asociado a la base de datos local.
- **Clasificador.** Las noticias obtenidas se clasifican en diferentes tópicos, lo cual facilitará el acceso temático a las mismas. Es importante destacar que la mayoría de las fuentes proveen una clasificación manual de la cual nuestro clasificador podrá valerse al momento de integrar las noticias descargadas por los recolectores.
- **Modelador del Contexto del Usuario.** Este módulo esta a cargo de modelar el contexto temático del usuario. Esto hará posible el acceso a noticias sensibles a la tarea en la que el usuario se encuentre inmerso. Algunos de los algoritmos a utilizar están descriptos en detalle en trabajos anteriores ([7],[13])
- **Ranker.** El ranker es el módulo encargado de ordenar las noticias de acuerdo a su orden de importancia o interés para el usuario. Es importante destacar que los criterios a tener en cuenta para hacer un ranking de noticias difiere de los utilizados para ordenar páginas web.
- **Interfaz.** Este módulo es el encargado de interactuar con el usuario. El mismo no sólo facilitará el acceso a las noticias, sino que también facilitará la generación de anotaciones y el agregado de preferencias por parte del usuario.

3. Línea de Investigación - Resultados Esperados

A partir de la arquitectura propuesta esperamos desarrollar una serie de mecanismos novedosos que facilitarán el manejo cualitativo del material almacenado en nuestras bases de datos. El objetivo de esta línea de investigación es desarrollar un sistema que sea capaz de manejar repositorios de noticias con contenidos estructurados e información semántica. El sistema propuesto permitirá generar recomendaciones basadas en conceptos, permitiendo que diversas aplicaciones puedan reutilizar los datos almacenados. Además, los componentes de nuestro sistema interactuarán entre sí mediante servicios web, lo cual facilitará la integración con agentes inteligentes.

La contribución de nuestra propuesta se centrará en el modelado del contexto temático del usuario, la incorporación de mecanismos de inferencia capaces de razonar sobre las preferencias del usuario, la utilización de anotaciones adicionales provistas por los lectores y el diseño y evaluación de algoritmos híbridos de rankings.

En el diseño de un algoritmo de ranking destinado a noticias debemos tener en cuenta varios aspectos, entre los cuales cabe destacarse los siguientes:

- Podemos esperar mucho menos spam que el encontrado usualmente en la web ya que las noticias provienen de fuentes controladas.
- Cada noticia publicada es un trozo de información reciente. Por lo tanto, es probable que nadie apunte a ella, lo que vuelve ineficaz el uso de los algoritmos clásicos basados en enlaces.
- Las fuentes de noticias proveen información continuada y periódica. En consecuencia esta información puede estar repetida en distintos periodos de tiempo o en diversas fuentes.

- El periodismo está basado en opiniones. Por tal motivo, dependiendo de la fuente consultada, es posible la inclusión, exclusión o jerarquización de información asociada a un mismo evento. En casos extremos, es también posible la manipulación y distorsión de la realidad. Por tal motivo, es necesario incorporar nociones de confiabilidad.
- Diversas noticias pueden estar relacionadas (tratar sobre un mismo tema) a pesar de que no exista un enlace explícito entre las mismas.
- Los medios están condicionados por el tiempo y la información provista por los mismos presentan sólo fragmentos de la realidad que podrían no estar actualizados. Además, los sitios de publicación de noticias pueden tener periodicidad variable (mensual, semanal, diario, horario) lo cual puede tener especial importancia al momento de ordenar e integrar noticias de varias fuentes.

Incorporar estos aspectos en el diseño de algoritmos de integración y ranking de noticias presenta nuevos desafíos de investigación que esperamos abordar.

Referencias

[1] http://newsbot.msnbc.msn.com/.

[2] http://news.google.com/.

[3] http://news.yahoo.com/.

[4] http://www.atomenabled.org/.

[5] http://www.nielsen-netratings.com/.

[6] http://www.rssboard.org/.

[7] CHESÑEVAR, C. I., LORENZETTI, C. M., MAGUITMAN, A. G., SAGUI, F. M., AND SIMARI, G. R. Exploiting user context and preferences for intelligent web search. In *Proceedings del 8vo Workshop de Investigadores en Ciencias de la Computación (WICC)* (Morón, Argentina, May 2006), Universidad de Morón, pp. 149–153.

[8] CHESÑEVAR, C. I., AND MAGUITMAN, A. G. Combining argumentation and web search technology: Towards a qualitative approach for ranking results. *Intl. Journal of Advanced Computational Intelligence 9*, 1 (2005), 53–60.

[9] DUNLAVY, D. D. QCS: A tool for querying, clustering, and summarizing documents.

[10] GABRILOVICH, E., DUMAIS, S., AND HORVITZ, E. Newsjunkie: Providing personalized newsfeeds via analysis of information novelty, 2004.

[11] GULLI, A. The anatomy of a news search engine. In *WWW 2005* (Chiba, Japan, May 2005).

[12] HENZINGER, M., CHANG, B., MILCH, B., AND BRIN, S. Query–free news search, 2003.

[13] LORENZETTI, C. M., SAGUI, F. M., MAGUITMAN, A. G., SIMARI, G. R., AND CHESÑEVAR, C. I. Incremental methods for context-based web retrieval. In *Proceedings del 12vo Congreso Argentino de Ciencias de la Computación (CACiC)* (San Luis, Argentina, Oct. 2006), Universidad Nacional de San Luis, pp. 1243–1254.

[14] RADEV, D. R., BLAIR-GOLDENSOHN, S., ZHANG, Z., AND RAGHAVAN, R. S. NewsInEssence: A system for domain-independent, real-time news clustering and multi-document summarization.

[15] REIS, D., GOLGHER, P., SILVA, A., AND LAENDER, A. Automatic web news extraction using tree edit distance, 2004.