Formalización de Web Mining como Conocimiento Estructurado

Gabriel R. Filocamo **Carlos I. Chesñevar**

Laboratorio de Investigación y Desarrollo en Inteligencia Artificial (LIDIA)
Departamento de Ciencias e Ingeniería de la Computación
Universidad Nacional del Sur
Av.Alem 1253 – B8000CPB Bahía Blanca – República Argentina
Tel/Fax: (+54) (291) 459 5135/5136 – Email: {grf,cic}@cs.uns.edu.ar
Palabras clave: Web Usage Mining, Data Mining, lenguajes de marcaje

Resumen

El descubrimiento de conocimiento a través de Web Mining involucra el proceso de recuperar datos de fuentes de textos disponibles en la Web tales como boletines de noticias, grupos de noticias, documentos HTML, base de datos, etc. Estos recursos poseen diversos formatos por lo cual antes de ser usados para la extracción de conocimiento necesitan algún tipo de procesamiento preliminar. Formalizar este procesamiento constituye un desafío por la gran diversidad de formatos existentes y la necesidad de evitar redundancias en la manipulación y representación de la información.

Este trabajo sintetiza los principales aspectos de una línea de investigación que se ha comenzado a desarrollar para abordar el problema antes planteado. El eje principal del acercamiento propuesto involucra la definición de un *lenguaje de marcaje* estandarizado que permita facilitar la tarea de web mining.

1 Introducción y motivaciones

La minería de datos de la Web (denominada comúnmente por su nombre en inglés *Web Mining*) es un área de las Ciencias de la Computación que involucra el uso de técnicas y acercamientos basados en la minería de datos (*Data Mining*) orientados al descubrimiento y extracción automática de información de documentos y servicios de la Web. El Web Mining ha despertado gran interés en la actualidad, particularmente debido a los avances de la comunidad científica en distintas líneas de investigación relacionadas con Data Mining orientado a la World Wide Web. En los últimos años esto se ha potenciado fuertemente en virtud del gran aumento en volumen del tráfico, tamaño y complejidad de las fuentes de información disponibles en la Web y el reciente interés en el desarrollo del comercio electrónico (e-commerce).

El descubrimiento de recursos a través de Web Mining es el proceso de recuperar datos (que pueden estar online o offline) de fuentes de textos disponibles en la Web tales como boletines de noticias, grupos de noticias, documentos HTML, base de datos, etc. Una vez que estos recursos han sido descubiertos comienza la etapa de selección y transformación de los datos originales recuperados en información procesada. Las transformaciones pueden consistir en remover palabras, cadenas, o realizar un pre-procesamiento que apunte a obtener una representación adecuada. Posteriormente se tiene una etapa de generalización donde típicamente son usadas técnicas de *Data Mining* y *Machine Learning*. Data Mining es un proceso de extracción no trivial de información implícita, previamente desconocida y potencialmente útil a partir de grandes volúmenes de datos (por ejemplo, para capturar el comportamiento de usuarios a partir de un archivo bitácora). Machine Learning involucra distintas técnicas y algoritmos de aprendizaje para automatizar la extracción de conocimiento a partir de grandes conjuntos de datos (como por ejemplo el uso de programas "spiders" en la Web).

Debe tenerse en cuenta que los seres humanos juegan un rol importante en el proceso de descubrimiento de conocimiento o información en la Web dado que ésta es un medio interactivo. Este hecho es importante al momento de validar e interpretar la información obtenida. Recientemente se han perfeccionado diversas técnicas para automatizar el proceso de Web Mining, recurriendo a técnicas avanzadas de indexación, algoritmos de búsqueda altamente eficientes y la utilización de agentes inteligentes para la paralelización de consultas y extracción de información.

Formalizar adecuadamente el proceso de webmining constituye un desafío por la gran diversidad de formatos existentes y la necesidad de evitar redundancias en la manipulación y representación de la información. Este trabajo sintetiza los principales aspectos de una línea de investigación que se ha comenzado a desarrollar para abordar este problema antes planteado. El eje principal del acercamiento propuesto involucra la definición de un *lenguaje de marcaje* estandarizado que permita facilitar la tarea de web mining.

2 Web Mining

La tarea de Data Mining es la búsqueda o extracción de conocimiento de grandes conjuntos de datos. Usualmente estos conjuntos están representados por bases de datos. Para descubrir conocimiento se usan diversas técnicas tales como machine learning, estadística y visualización. La aplicación de estas técnicas para descubrir y extraer automáticamente información de documentos y servicios de la Web se conoce como *Web Mining* [Etz96]. La Web puede ser vista como un gran conjunto de datos que en general se define como una enorme, diversa y dinámica colección de documentos hipertextuales entrelazados. En consecuencia podemos aplicar a la Web las técnicas de Data Mining para la búsqueda y extracción de conocimiento. Entre las más destacadas podemos mencionar:

Clasificación [HCC93] El descubrimiento de reglas de clasificación permite desarrollar un perfil de items que pertenecen a un grupo particular de acuerdo a los atributos que tienen en común. Este perfil puede luego ser usado para clasificar nuevos items de datos que son añadidos a la base de datos. En el caso particular de Web Mining, las técnicas de clasificación permiten desarrollar un perfil de clientes quienes acceden a archivos particulares de un server basado en sus patrones de acceso.

Reglas de Asociación [RAS93] Estas proveen un mecanismo útil para descubrir correlaciones entre items que pertenecen a transacciones de clientes en una base de datos comercial. En la Web, este problema equivale a descubrir las correlaciones entre referencias a diversos archivos disponibles en un server por un cliente dado. Cada transacción esta constituida por un conjunto de URLs accedidos por un cliente en una visita a un server. Descubrir este tipo de reglas puede ayudar a una organización en el comercio electrónico a desarrollar estrategias de marketing mas efectivas. Además, reglas descubiertas de bitácoras de accesos Web (Web access logs) pueden proveer una mejor indicación de como organizar un sitio Web.

Episodios Frecuentes [MTV95] Esta técnica trata de encontrar patrones entre transacciones tales que la presencia de un conjunto de items es seguido por otro item en el conjunto de transacciones ordenadas por estampillas de tiempo. En general, en los server Web, la visita de un cliente es registrada por un período de tiempo. El descubrimiento de patrones secuenciales en bitácoras de accesos a un server Web (Web server access logs)

permite a organizaciones predecir patrones de navegación de usuarios y ayudar en futuras estrategias de marketing, por ejemplo, a que grupos dirigir las diversas ofertas y promociones. Por el análisis de esta información, Web Mining puede determinar relaciones temporales entre items de datos.

Generalmente, Web Mining se divide en función a los datos analizados en tres categorías:

- **Minería de Web por Contenidos** (o *Web Content Mining*): Es el proceso de extraer conocimiento de los contenidos de los documentos o sus descripciones.
- **Minería de Web por Estructura** (o *Web Structure Mining*): Es el proceso de inferir conocimiento de la organización de la Web.
- **Minería de Web por Uso** (o *Web Usage Mining*): Es el proceso de extraer patrones interesantes a partir de archivos bitácora (log files) que contengan información sobre el comportamiento de un usuario en relación a su interacción con la Web.

Este proceso de Web Mining no es trivial por lo cual podemos identificar distintas subtareas: a) *descubrimiento de recursos* (localizar documentos no conocidos y servicios en la Web); b) *extracción de información* (identificar y extraer automáticamente información específica de la Web); c) *generalización* (luego de automatizar el descubrimiento y extracción de información, el próximo paso es generalizar estas experiencias, tanto para un sitio particular como también para múltiples sitios); d) *análisis y validación* (validación de los patrones extraídos).

3 Lenguajes de marcaje[1] y su aplicación en webmining

La necesidad de estructurar documentos de una manera estándar para facilitar el intercambio y manipulación de datos dio lugar ya en la década del 60 al desarrollo de un lenguaje de marcaje general estándar llamado SGML *(Standard Generalized Markup Language)*. Este lenguaje de marcado era muy poderoso, su complejidad obligó al desarrollo de alternativas más manejable. Así en 1991 surge el lenguaje HTML (*Hyper Text Markup Language*), definido como un subconjunto de SGML. Este lenguaje se difundió altamente con la evolución de Internet al constituirse en el lenguaje de base para desarrollo de páginas web.

El lenguaje HTML brinda un conjunto de marcas o 'tags' que pueden usarse para estructurar un documento hipertextual. Pese a su expresividad, HTML presenta significativas limitaciones por su conjunto fijo de etiquetas y limitada capacidad para representar información estructurada. Recientemente comenzó a popularizarse el lenguaje XML (*eXtended Markup Language*) [XML], el cual es también un subconjunto de SGML con una simplicidad similar a la de HTML pero con el beneficio de ser extensible y estructurado. XML provee una especificación para el diseño de una *familia* de lenguajes de marcado, definiendo un formato de texto estandarizado para la representación de información estructurada en la Web.

La evolución reciente de XML dio lugar a la aparición de un gran número de lenguajes de marcaje para dominios especificos tales como matemática, química, y otros. Se han iniciado líneas de desarrollo en tal sentido en el área de Web Usage Mining a través del lenguaje LOGML (*Log Markup Language*) [PKZ01]. LOGML está desarrollado a partir de XML 1.0, y permite describir reportes de *logs* provenientes de servidores web. LOGML ha sido diseñado

[1] Utilizaremos el término *lenguaje de marcaje* como traducción de su correspondiente en inglés *mark-up language*.

para facilitar el proceso de Web Mining, además de almacenar información detallada extraída a partir de los weblogs. El uso de documentos generados via LOGML simplifica considerablemente los pasos de preprocesamiento requeridos para web usage mining. Cabe señalar que el desarrollo de lenguajes de marcaje para Web Mining es aun un tema en desarrollo en la actualidad.

4 Investigación en desarrollo

En el ámbito del LIDIA hemos estado trabajando en el desarrollo de una arquitectura para Web Mining que permita la representación y manipulación de información estructurada. Entendemos que la representación estructurada de la información juega un rol vital en el proceso de Web Mining, y para esto se requiere contar con lenguajes de marcaje que provean la ontología apropiada para abordar dos tareas principales:

- La representación de información de manera estructurada;
- La posibilidad de realizar inferencias a partir de esa información, a través de un motor de inferencias basado en *reglas*.

En este último aspecto, cabe destacar que la representación de distintos tipos de *reglas* a través de lenguajes de marcaje en la Web se ha vuelto un tema de especial relevancia. La formalización adecuada de reglas de inferencia fue identificada como un aspecto de diseño (Design Issue) en el diseño de la Web semántica[2] por su importancia en relación a los sistemas basados en conocimiento y las aplicaciones basadas en agentes inteligentes. Iniciativas recientes motivaron el desarrollo de un lenguaje denominado RuleML (*Rule Markup Language*), que permite representar reglas en XML para realizar razonamiento hacia atrás, hacia adelante, reescritura de información, y otras tareas transformacionales.

Nuestra línea de investigación tuvo su origen en el estudio de *reglas de asociación* en el contexto de bases de datos transaccionales [FG02]. Este problema permitió identificar el rol de los lenguajes de marcaje para la representación de reglas de asociación en distintos contextos. Posteriormente se abordó el problema de formalizar apropiadamente el proceso de Web Mining. Se estudiaron entonces los lenguajes de marcaje existentes para tal fin (ej. LOGML), y se abordaron posibles mejoras y extensiones a fines de considerar no solo *logs* sino otros tipos de información más generales.

En este contexto se buscaron integrar los aspectos de representación de información y de mecanismos de inferencias a través de reglas. En tal sentido, en el ámbito del LIDIA se cuenta con un sólido desarrollo en el área de razonamiento con información incompleta utilizando *argumentación rebatible* [SL92, CML00]. Contar con un lenguaje apropiado para abordar el problema de Web Mining permitiría integrar una base de conocimiento (expresada en dicho lenguaje) con una máquina de inferencia argumentativa, permitiendo automatizar el proceso de toma de decisiones basadas en resultados de algoritmos de Web Mining.

5 Conclusiones

Formalizar apropiadamente el proceso de Web Mining es una tarea ardua y compleja. Nuestro enfoque para abordarla consiste en establecer un lenguaje de representación de conocimiento

[2] Ver `http://www.w3.org/DesignIssues/Rules.html`.

flexible pero suficientemente expresivo que permita capturar los principales aspectos asociados a este proceso. Inicialmente nos hemos focalizado en Web Usage Mining, con la intención de extendernos posteriormente hacia otras variantes de Web Mining. Cabe destacar que en Web Usage Mining ya se han desarrollado arquitecturas específicas, tales como WebSIFT [CSM97] y WUM [SF98].

Los desarrollos logrados en el área de Web Mining son extremadamente recientes. En tal sentido, entendemos que la integración de técnicas de representación de conocimiento (en nuestro caso, a través de lenguajes de marcaje apropiados) con el uso de reglas para razonar a partir de esta información es un área de central interés, posibilitada en gran medida por la aparición de XML como estándar de representación, y sus subsecuentes iniciativas (como RuleML y LOGML). Los resultados obtenidos hasta el momento han sido promisorios, restando aún considerar en detalle la integración de una base de conocimiento con un motor de inferencia apropiado (ej. uno basado en argumentación). Parte de nuestra investigación actual se centra en este tema.

Referencias

[CML00] Chesñevar, C. I., Maguitman, A., and Loui, R. Logical Models of Argument. *ACM Computing Surveys 32*, 4 (December 2000), 337–383.

[Coo00] Cooley, R. Web usage mining: Discovery and application of interesting patterns from web data, 2000.

[CSM97] COOLEY, R., SRIVASTAVA, J., and MOBASHER, B. Web mining: Information and pattern discovery on the world wide web, 1997.

[Etz96] Etzioni, O. The world-wide web: Quagmire or gold mine? *Communications of the ACM 39*, 11 (1996), 65–68.

[FG02] Filocamo, G. R., and Grandinetti, W. Reglas de asociación para datamining – teoría y aplicaciones. *Tesis de Licenciatura – Dep. de Cs. e Ing. de la Computación – Universidad Nacional del Sur* (2002).

[HCC93] Han, J. W., Cai, Y. D., and Cercone, N. Data-driven discovery of quantitative rules in relational databases. *Ieee Trans. On Knowledge And Data Engineering 5* (Feburary 1993), 29–40.

[KB00] Kosala, and Blockeel. Web mining research: A survey. *SIGKDD: SIGKDD Explorations: Newsletter of the SIG on Knowledge Discovery and Data Mining, ACM 2* (2000).

[MTV95] Mannila, H., Toivonen, H., and Verkamo, A. I. Discovering Frequent Episodes in Sequences. In *Proc. of the 1st International Conference on Knowledge Discovery and Data Mining (KDD-95)* (Montreal, Canada, August 1995), U. M. Fayyad and R. Uthurusamy, Eds., AAAI Press.

[PKZ01] Punin, J. R., Krishnamoorthy, M. S., and Zaki, M. J. LOGML - XML language for web usage mining. In *WWW Posters* (2001).

[RAS93] Rakesh Agrawal, T. I., and Swami, A. Mining association rules between sets of items in large databases. In *SIGMOD-93* (May 1993), pp. 207–216.

[SCDT00] Srivastava, J., Cooley, R., Deshpande, M., and Tan, P.-N. Web usage mining: Discovery and applications of usage patterns from web data. *SIGKDD Explorations 1*, 2 (2000), 12–23.

[SF98] Spiliopoulou, M., and Faulstich, L. C. WUM: a Web Utilization Miner. In *Workshop on the Web and Data Bases (WebDB98)* (1998), pp. 109–115.

[SL92] Simari, G. R., and Loui, R. P. A Mathematical Treatment of Defeasible Reasoning and its Implementation. *Artificial Intelligence 53* (1992), 125–157.

[XML] Xml: Extensible markup language, http://www.w3.org/xml.

[ZXH98] Zaiane, O. R., Xin, M., and Han, J. Discovering web access patterns and trends by applying OLAP and data mining technology on web logs. In *Advances in Digital Libraries* (1998), pp. 19–29.