La semántica en la Web

Sandra ROGER
sroger@uncoma.edu.ar

Departamento de Ciencias de la Computación.
Universidad Nacional del Comahue.
Buenos Aires 1400, CP 8300, Neuquén, Argentina.
FAX: (54)(0299)4490313.

1. Introducción

En la web contamos con una fuente incalculable de textos de todo tipo. Lamentablemente, ésta presenta un crecimiento caótico de recursos, redundancia, posee una semántica implícita y cuenta con una falta de orden y de organización.

Según la literatura, a la web la podemos clasificar en:

- Estática.
- Dinámica (páginas que se generan producto de un clic o de alguna consulta).
- Oculta (invisible o profunda).
- Pública.
- Semántica (representada principalmente por metadatos).
- Indizable (la que encuentran los buscadores).

El porcentaje de la web oculta varía según los autores ([BY04], [Tur03], [Ouf01], [Ber00], [SP01]) debido a los métodos utilizados para obtener las cifras que relacionan la web oculta (invisible o profunda) y la indizable (visible o superficial). En cualquier caso, la primera aventaja considerablemente a la segunda.

Ricardo Baeza-Yates [BY04] afirma que la web estática supera los cuatro millones de páginas mientras que las dinámicas está en relación de 100 a 1 con las estáticas

La web semántica actualmente, tal como se pretende, no existe, el 5 % (según [BY04]) de la web actual está representada principalmente a través de metadatos, los cuales muchas veces contienen información no fideligna.

El propósito de este trabajo es presentar la motivación de nuestra línea de investigación, metas y desarrollos futuros. Fundamentalmente, nuestro interés se centrará en poder capturar el potencial de las técnicas y paradigmas basadas en conocimiento semántico para la representación de conocimiento, la localización, compartición e integración de recursos a mediante la WWW.

2. Motivación

Los problemas de la web actual son los siguientes:

- Se cuenta con una gran cantidad de texto el cual es interpretado por las máquinas como simples caracteres y enlaces a otras páginas.
- Los buscadores basan su consultas en palabras claves pudiendo recuperar información irrelevante y la información que necesitamos puede no ser devuelta a través de éstas búsquedas.
- La clasificación de si un texto es relevante o no es realizado por el usuario.
- No se cuenta con una buena representacion computable de la semántica involucrada en las utilidades de sitios web adaptativos, las cuales permiten la reconfiguración dinámica de acuerdo al perfil del usuario u otros aspectos relevantes.

Existen dos formas complementarias de tratar estos problemas. Una es a través de la incorporación de información adicional mediante anotaciones que provean semántica y así pueda ser interpretado por las máquinas. La otra forma, es extraer la información semántica de los recursos disponibles en la web por medio de programas tales como filtros, wrappers y programas de extracción específicos.

El objetivo de la web semántica es poder transformar la información disponible en conocimiento. Un ejemplo sencillo es si deseamos buscar información relacionada con "caballos", el buscador pueda recuperar información sobre "yeguas"también.

La definición según el sitio líder oficial de la web semántica (W3C)[1] no dice nada acerca de lo que es. Aunque otros autores (ej. [Cod03]) brindan una definición algo más específica de lo que pretende ser.

Actualmente, la única semántica contenida en la web es principalmente producto de los metadatos de las páginas. Muchas de las cuales se consideran "spamming de metadados"([BY04]) dado que pueden contener información no confiable, por ejemplo, llenar campos de información semántica como autor con datos ficticios o el típico "xxx".

El problema para hacer viable la introducción de la semántica en la web es alcanzar un entendimiento entre los usuarios, desarrolladores y tecnología: acordar estándares, realización de herramientas y tecnologías y la adopción de éstas por parte del mercado.

Como enumera Pablo Castells ([Cas02]) los campos de aplicación donde esta idea puede tener utilidad son:

- Comercio electrónico
- Gestión del conocimiento corporativo (intranet)
- Procesamiento de Lenguaje Natural
- Búsqueda de información en la web
- Enseñanza (recursos educativos reutilizables)

[1] World Wide Web Consortium

- Librerías digitales
- Turismo, Oscio, Patrimonio Cultural

Los investigadores están abocando sus esfuerzos al desarrollo de infraestructuras necesarias para la puesta en marcha de la web semántica y aplicaciones que puedan demostrar que es una meta realista y justifique su beneficio, a la vez que motiven al desarrollo y consumo de esta infraestructura.

Las principales líneas de interés de los investigadores son:

- Infraestructura
  - Lenguajes de definición de ontologías
  - Metodologías de desarrollo de ontologías
  - Desarrollo de vocabularios en dominios concretos
  - Servicios Web
- Problemas
  - Agentes
  - Visualización, navegación, gestión del diálogo
  - Aprendizaje de ontologías
  - Integración de ontologias
  - Aplicaciones

3. Conclusiones y Trabajos Futuros

La web semántica es una visión, actualmente se ha comenzado con la realización de herramientas apropiadas de estándares, pero aún está en sus primeros pasos.

Esta nueva idea impulsada por Tim Berners-Lee proporcionará un grado cualitativo sobre el potencial de la web, brindando, entre otros beneficios:

- Un desarrollo de aplicaciones con esquemas comunes.
- Una búsqueda a partir de inferencias.
- El incentivo de las transacciones entre empresas por medio del correo electrónico dado que la seguridad, unido a las firmas digitales serán facilitadas por medio de la semántica en la web.

Para poder lograr estas metas es necesario

- Establecer consenso entre los desarrolladores de software, aplicaciones y usuarios.
- Un lenguaje común basado en web, para establecer tipos de relaciones (utilizando lenguajes adecuados, para definir relaciones, y vocabularios y reglas que permitan definir las relaciones entre unos significados y otros).

- Agentes inteligentes y aplicaciones que exploten la semántica, expuestas mediante anotaciones.

Nuestro interés estará en poder realizar una investigación, análisis y comparación de las técnicas y paradigmas basadas en conocimiento semántico actuales para la representación de conocimiento, la localización, compartición e integración de recursos mediante la WWW, para poder capturar el potencial de las mismas.

Referencias

[Ber00] Michael K. Bergman. The deep web: Surfacing hidden value. *Bright Planet*, 2000.

[BY04] Ricardo Baeza-Yates. Excavando la web. *El profesional de la información*, 13(1), February 2004.

[Cas02] Pablo Castells. Aplicación de técnicas de la web semántica. *Workshop de investigación en entornos de interacción colectiva (COLINE'02). Granada*, November 2002.

[Cod03] Luis Codina. Internet invisible y web semántica: ¿el futuro de los sistemas de información en línea? *Revista Tradumática. http://www.fti.uab.es/tradumatica/revista*, (2), November 2003.

[Ouf01] Rehib Ouf. Le dynamisme du world wide web: Taille, croissance, visibilité, distribution et accessibilité de l'information. *Lyon, France: Ecole Nationale Supérieure des Sciences de l'Information et des Bibliothèques*, 2001.

[SP01] Chris Sherman and Gary Price. The invisible web. *Searcher*, 8(9):62–74, 2001.

[Tur03] Laura Turner. Doing it deeper: The deep web. *http://www.bhsu.edu/education/edfaculty/lturner/The Deep Web article1.doc*, 2003.