Universidad Nacional de Luján

Licenciatura en Sistemas de Información

Enriquecimiento Automático de Textos

Autor: Mariano Felice

Director: Mg. Gabriel H. Tolosa

Co-Director: Mg. Fernando R. A. Bordignon

~ 2009 ~

A mis padres y hermanos.

Al resto de mi familia.

Agradecimientos

En primer lugar, quiero agradecer a mis padres por enseñarme el valor de la educación y brindarme los medios para desarrollarme como estudiante y como persona. También a mis hermanos, quienes me acompañaron cotidianamente en mis éxitos y fracasos, y al resto de mi familia por su constante apoyo.

Quiero agradecer además a Santiago "Pipo" Banchero, un amigo y compañero con el que cursé toda mi carrera y compartí muchas cosas, académicas y personales. Estudiar no habría sido lo mismo sin su compañía.

Agradezco especialmente a mi director, Gabriel H. Tolosa, por haber sido mi guía en esta última etapa y ser un ejemplo de profesionalismo lejos de la soberbia, siempre dispuesto a darme una mano durante toda mi carrera. Del mismo modo, quiero extender las gracias a mi codirector, Fernando R. A. Bordignon, por sus consejos y motivación tanto en este como en otros emprendimientos.

También agradezco a Luis Aguado por su incentivo y revisiones desinteresadas de este trabajo, y a Juan Pablo Sarubbi, por haberme ayudado con la implementación del prototipo.

Finalmente, pero no menos importante, a la Universidad Nacional de Luján, por haberme brindado la posibilidad de estudiar gratuitamente y conocer gente muy valiosa e interesante durante el transcurso de mis estudios.

Resumen

Desde el origen de la humanidad, la información ha sido siempre un recurso codiciado cuyo valor se ha incrementado con el paso del tiempo. Por ello, se ha puesto mucho esfuerzo en el desarrollo y evolución de tecnologías para el manejo de información, abarcando su producción, publicación y distribución. Al día de hoy, los materiales escritos como libros, revistas y monografías siguen siendo una de las principales vías para la transmisión de conocimientos y gozan de amplia aceptación. Sin embargo, con el surgimiento de Internet y la World Wide Web, el acceso a la información cobró nuevas dimensiones, modificando, incluso, la forma de leer. La introducción del hipertexto como una nueva forma de presentación de información permitió superar las limitaciones impuestas por los materiales físicos como el papel, que impone una lectura estrictamente lineal.

A diferencia de textos escritos en medios físicos, el hipertexto permite la vinculación de elementos del texto con otros fragmentos, documentos o recursos multimedia de manera de poder "saltar"de un punto a otro a lo largo de la lectura. Este mecanismo junto con la posibilidad de incorporar elementos no textuales (como imágenes, audio y video) modifica la experiencia de leer y sienta las bases para el desarrollo de sistemas de información más sofisticados. Actualmente, la Web incorpora una amplia variedad de servicios donde los usuarios pueden participar de forma activa aportando recursos a la red (Web 2.0). Entre ellos, son populares los blogs, sitios colaborativos como Wikipedia y aplicaciones con prestaciones similares a las de escritorio. La Web se ha transformado en un servicio ubicuo, siendo accesible tanto desde una computadora como desde dispositivos móviles. Esto les ha permitido a los usuarios acceder a información en situaciones en las que requieren conocimientos específicos, por lo general de manera rápida y concisa. Concretamente, todos estos nuevos servicios de información son de

gran utilidad para la lectura ya que podrían ser utilizados para complementar textos con recursos relacionados y brindar información específica sobre elementos particulares.

En consecuencia, se presenta el siguiente trabajo como la primera aproximación al Enriquecimiento Automático de Textos, una tarea que tiene por objetivo complementar un texto con recursos de la Web en forma automática. Este enriquecimiento, que carece de precedentes formales, está orientado a la transformación de textos planos lineales en hipertextos que brinden información contextual y recursos multimedia sin esfuerzo por parte de los usuarios. De esta forma, los lectores pueden transformar textos en hipertextos auto-explicativos con el objetivo de lograr una mayor comprensión y evitar búsquedas individuales de información afín.

El Enriquecimiento Automático de Textos presentado en este trabajo combina técnicas de de las áreas de Recuperación de Información (IR) y Extracción de Información (IE). Concretamente, se presenta una arquitectura y metodología compuesta por tres tareas principales:

1. *Reconocimiento de Entidades:* Es una subtarea de IE de amplia difusión cuyo objetivo es la identificación de elementos clasificables de acuerdo a alguna taxonomía previamente definida. Dichos elementos se denominan *entidades* y representan generalmente lugares, organizaciones y personas. En este trabajo no sólo se propone una clasificación de entidades extendida sino que además se incursiona en nuevas técnicas de reconocimiento.

2. *Enriquecimiento de Entidades:* Es la tarea encargada de efectuar caracterizaciones de entidades mediante la obtención de recursos apropiados de la World Wide Web. Si bien existen algunos trabajos con objetivos similares, este es el primero en formalizar sus objetivos y proponer una metodología.

3. *Presentación de Resultados:* Es la tarea orientada a la presentación óptima de los textos enriquecidos, valiéndose de áreas como la Visualización de Información y la Interacción Hombre-Máquina.

Para el desarrollo de este trabajo, también se ha hecho uso de técnicas de Procesamiento de Lenguaje Natural como la desambiguación, la cual es aplicada para la clasificación y obtención de recursos de entidades.

El enriquecimiento de textos descripto se propone específicamente para textos en español, aunque puede ser adaptado a otros idiomas si se procuran ciertas modificaciones en su implementación. Adicionalmente, los enriquecimientos obtenidos con esta metodología son de carácter primordialmente informativo, en estrecha relación con los objetivos perseguidos. Sin embargo, al tratarse de una arquitectura genérica, es posible adaptar sus prestaciones a distintos objetivos, que van desde lo educativo a lo comercial.

Aportes

En este trabajo se incluyen varios aportes al área de Recuperación de Información y Extracción de Información, en su mayoría como consecuencia del desarrollo de soluciones a problemas encontrados. Entre los aportes más significativos se incluyen:

- la primera caracterización y especificación formal del concepto de "Enriquecimiento Automático de Textos",

- nuevos lineamientos para el Reconocimiento de Entidades, entre los que se incluyen la adopción de *gazetteers* de actualización dinámica y la reutilización de entidades resueltas por correferencia,

- nuevas técnicas para la clasificación y desambiguación de entidades basada en consultas a servicios web como motores de búsqueda,
- la primera arquitectura y metodología para el Enriquecimiento de Entidades,
- la *verificación de identidad*, una solución simple para desambiguar entidades durante la extracción de recursos,
- la aplicación de la metodología a textos en español.

Este trabajo no sólo presenta el Enriquecimiento Automático de Textos de manera formal sino que propone su aplicación al idioma español, sentando las bases para futuros trabajos en el área.

Organización del Trabajo

El Capítulo 1 brinda una introducción al hipertexto y la World Wide Web, componentes base del enriquecimiento de textos. El Capítulo 2 describe los distintos enfoques para el Reconocimiento de Entidades mientras que en el Capítulo 3 se comentan los distintos servicios que se aproximan a la noción de Enriquecimiento de Entidades. Los siguientes capítulos, por el contrario, describen en detalle las propuestas de este trabajo. En el Capítulo 4 se brinda una caracterización del Enriquecimiento Automático de Textos y su arquitectura, además de proponerse una metodología de Reconocimiento de Entidades y formalizarse la tarea de Enriquecimiento de Entidades. El Capítulo 5 describe la implementación de un prototipo para la validación de la propuesta y el Capítulo 6 concluye el trabajo con algunas consideraciones finales y lineamientos para trabajos futuros.

Tabla de Contenidos

Capítulo 1

Introducción

1.1. Orígenes y Evolución del Hipertexto

Mucho antes del desarrollo de las computadoras personales, y sin habérselo propuesto, el ingeniero estadounidense Vannevar Bush sentaba las bases de lo que posteriormente se conocería como hipertexto. En su artículo *As We May Think* [Bush, 1945], publicado en The Atlantic Monthly en julio de 1945, Bush criticaba los métodos de gestión de información utilizados en aquella época, haciendo hincapié en la ineficiencia que imponía la revisión secuencial de grandes volúmenes de información. Como solución, proponía un dispositivo electrónico de consulta, bautizado Memex, donde los usuarios pudieran almacenar toda la información que desearan utilizando microfilms, un soporte en auge por aquellos años. Este sistema imaginario permitiría generar referencias cruzadas entre los distintos documentos almacenados y consultarlos de forma rápida y eficiente, pudiendo ver en cada momento qué documentos contenían información relacionada con el documento consultado. En su visión, Bush introducía un nuevo modelo de almacenamiento y recuperación de información basada en la asociación, acercándose de este modo a la forma natural en que los humanos gestionamos la información.

Aunque este dispositivo nunca llegó a implementarse, sí sentó las bases para que otros investigadores continuaran explorando nuevas formas de relacionar y recuperar información.

El primero de ellos, el filósofo y pionero Theodor Holm Nelson, había comenzado en 1960 a desarrollar un prototipo de procesador de textos con el objetivo de vincular toda

la literatura universal en un único repositorio. Este sistema englobaría una red mundial de documentos totalmente interrelacionados mientras que brindaría funciones de control de versiones y derechos de autor. Cinco años más tarde, Nelson formalizó su propuesta en un artículo presentado ante la Association for Computer Machinery[1] (ACM), donde definió el término "hipertexto" como *"un cuerpo de material escrito o pictórico interconectado en una forma compleja que no puede ser representado en forma conveniente haciendo uso de papel"* [Nelson, 1965]. Posteriormente, en el año 1967, Nelson bautizaría su prototipo como Proyecto Xanadú, continuando su desarrollo hasta nuestros días, en pugna con la World Wide Web y sin verdadero éxito.

El concepto inicial de hipertexto fue refinado posteriormente por Nelson en su libro *Literary Machines* [Nelson, 1980], donde aclaró: *"Con 'hipertexto' me refiero a una escritura no secuencial, a un texto que bifurca, que permite que el lector elija y que se lea mejor en una pantalla interactiva. De acuerdo con la noción popular, se trata de una serie de bloques de texto conectados entre sí por nexos, que forman diferentes itinerarios para el usuario. (…) Hipertexto es una combinación de textos en lenguaje natural y la capacidad de la computadora de exposición dinámica de un texto no lineal"*.

Sin embargo, Nelson no era el único investigador abocado a aplicar las ideas de Bush. Ya en 1962, el investigador Douglas C. Engelbart había expuesto su iniciativa para crear un avanzado sistema de información orientado a la comunidad científica basado en la integración de "medios para potenciar el intelecto humano" [Engelbart, 1962]. En aquel trabajo, mencionaba: *"La mayoría de las formas de estructuración que les mostraré surgen de la simple capacidad de establecer vínculos arbitrarios entre distintas subestructuras y ordenarle posteriormente a la computadora que muestre un conjunto de subestructuras vinculadas con cualquier posición relativa que designemos entre las distintas subestructuras. Pueden designarse tantas clases de vínculos distintos como se desee, de manera que puedan especificarse distintas presentaciones y tratamientos para*

[1] http://www.acm.org

cada tipo"[2]. Claramente, Engelbart señalaba la necesidad de conexión entre documentos. Ese mismo año, comenzaría a diseñar e implementar oN-Line System (NLS) [Engelbart, 1968], un sistema colaborativo de edición, almacenamiento y búsqueda de publicaciones científicas, basado fundamentalmente en su trabajo de investigación. La versión funcional de NLS, presentada al público en San Francisco en 1968, podría considerarse hoy en día como el primer sistema de hipertexto operativo desarrollado.

Por su parte, y desde 1967, Ted Nelson había comenzado a trabajar junto a Andries Van Dam en el desarrollo del Hypertext Editing System, el primer sistema de edición de hipertextos concebido bajo ese concepto.

Con el advenimiento de nuevas tecnologías, el hipertexto comenzó a afianzarse y dar paso a refinamientos y mejoras. Una de las contribuciones más importantes del área la constituiría ENQUIRE [Berners-Lee, 1980], un sistema para notebooks creado en 1980 por Timothy Berners-Lee dentro del Conseil Européen pour la Recherche Nucléaire[3] (CERN) de Ginebra, Suiza. Este sistema, precursor de la World Wide Web, permitía almacenar y relacionar información mediante enlaces entre documentos, y era utilizado por Berners-Lee para organizar información importante de manera similar a un índice.

Nuevos sistemas de edición de hipertexto comenzaron a desarrollarse y popularizarse durante los 80, entre ellos HyperCard [Atkinson, 1987] para Macintosh en 1987, impulsando así el desarrollo para distintas plataformas.

Hacia finales de la década, el surgimiento y popularización de la tecnología multimedia extendió los límites del hipertexto, al permitir la incorporación de imágenes, videos y sonidos que podían ser vinculados de igual manera que los textos. Es decir, un

[2] Texto original: "*Most of the structuring forms I'll show you stem from the simple capability of being able to establish arbitrary linkages between different substructures, and of directing the computer subsequently to display a set of linked substructures with any relative positioning we might designate among the different substructures. You can designate as many different kinds of links as you wish, so that you can specify different display or manipulative treatment for the different types."*

[3] http://www.cern.ch

documento podía ahora incluir enlaces a archivos de audio, video o imágenes además de textos. Así, se estableció el término "hipermedia" para referirse a los hipertextos multimedia, aunque no sería tan difundido.

En la década del 90, Internet, como medio de conexión entre computadores geográficamente distantes, abriría nuevas posibilidades para el crecimiento del hipertexto.

La conexión del CERN a la nueva red sumada al constante impulso que adquiría el hipertexto y las experiencias previas con ENQUIRE, incitaron a Berners-Lee a elaborar un nuevo proyecto. Esta vez, el objetivo era permitir que miembros del CERN conectados a la red pudieran relacionar y compartir información en forma de hipertextos [Berners-Lee, 1989]. En cierto modo, era una reformulación del Proyecto Xanadú propuesto por Nelson décadas atrás, aunque claramente más práctico y menos ambicioso. En 1989, Berners-Lee presentó su propuesta al CERN, la cual fue aceptada y posteriormente implementada en colaboración con Robert Calliau. Para 1990, el proyecto había adoptado el nombre de "World Wide Web", a diferencia de "WorldWideWeb" (sin espacios), el primer editor y navegador web creado por el equipo. La utilización de este navegador hizo necesario definir un lenguaje especial de escritura de hipertexto por lo cual se formalizó una primera especificación de lo que más tarde se constituiría como Lenguaje de Marcado de Hipertexto, o HTML[4] (del inglés, HyperText Markup Language).

En los años subsiguientes, la evolución de las tecnologías de hipertexto adquirió mayor fuerza e interés, dando nacimiento a navegadores web más sofisticados y poniendo a la World Wide Web, o simplemente Web, a disponibilidad de todo el mundo. La Web dejaba de ser un simple "conjunto de nodos interconectados por enlaces" [W3C, 1995] para convertirse en el mayor sistema de información distribuido.

[4] http://www.w3.org/TR/html401/

Gracias a su amplio impacto y aceptación, la Web ha logrado consolidarse como uno de los mayores repositorios de información y servicios disponibles en la actualidad, acompañada fuertemente por el crecimiento de Internet. Según la organización Internet World Stats, Internet ha experimentado un crecimiento sostenido y exponencial desde su creación, alcanzando 1.565 millones de usuarios en diciembre de 2008, un equivalente al 23,3% de la población mundial[5].

Este incesante crecimiento de Internet ha potenciado la utilización de la Web en todo el mundo, no sólo como plataforma de consulta de hipertextos sino como un nuevo medio para actividades tan variadas como el comercio, la mensajería y el entretenimiento. Servicios como el correo electrónico y el chat comenzaron a brindar interfaces web, inaugurando una tendencia que se extendería ampliamente. Los diversos intereses comenzaban a hacerse un lugar en la web: la publicidad, la compra-venta de productos, la transmisión de programas de radio y televisión, la difusión de noticias, las operaciones financieras y un sinnúmero de otras actividades habituales.

A partir del año 2004, comenzó a proliferar en la World Wide Web una nueva ola de servicios orientados a la participación activa de los usuarios. Surgen así los foros de discusión, blogs (sitios personales donde se exponen experiencias y opiniones), wikis (sitios web educativos y colaborativos), fotologs (blogs con álbumes de fotos), redes sociales (agrupaciones de usuarios vinculados por un interés común), folcsonomías (clasificaciones colaborativas de contenido), sitios de videos personales, la sindicación de contenido y otras innovadoras prestaciones.

Con estos cambios, se desdibuja la diferencia entre los publicadores de contenido y los usuarios finales; ahora son los usuarios quienes generan su propio contenido. Este fenómeno de colaboración y participación en el espacio web inauguró una nueva generación: la Web 2.0. Entre los grandes exponentes de esta nueva Web, pueden

[5] http://www.internetworldstats.com/emarketing.htm. Consultado el 09 de febrero de 2009.

mencionarse sitios de gran popularidad como YouTube[6], Flickr[7], del.icio.us[8] y Wikipedia[9], una enciclopedia colaborativa libre y de calidad comparable con la tradicional Enciclopedia Britannica [Giles, 2005].

La popularidad de la World Wide Web no es casual, sino que ha sido construida en base a su fácil acceso, variedad y disponibilidad. Todo recurso incluido en la Web puede ser ofrecido o accedido de manera remota y simultánea por cualquier equipo conectado a ella, lo que la transforma en una red de información global de gran escalabilidad. No sorprende entonces que la sociedad se haya volcado cada vez más a la Web para buscar y aportar información hasta transformarla en su fuente de consulta predilecta [Estabrook, 2007].

1.2. Caracterización del Hipertexto

Como se describió con anterioridad, el hipertexto evolucionó notablemente desde los días de Nelson por lo que varios críticos e investigadores intentaron redefinir el concepto. Históricamente, el término hipertexto se utilizó de manera indistinta con tres significados diferentes:

1. un modelo de organización no secuencial de información,
2. los sistemas de creación de documentos hipertextuales y
3. los documentos hipertextuales en sí mismos.

Para evitar confusiones y establecer un vocabulario adecuado, los investigadores optaron por utilizar los términos “hipertexto”, “sistema de gestión de hipertextos” e “hiperdocumento” para cada uno de los conceptos mencionados, aunque el uso común

[6] http://www.youtube.com
[7] http://www.flickr.com
[8] http://del.icio.us
[9] http://www.wikipedia.org

derivó en la popularización de "hipertexto" como sinónimo de "hiperdocumento" [Lamarca Lapuente, 2007].

Resulta interesante explorar en detalle las diversas reflexiones sobre el hipertexto a lo largo de su historia. El mismísimo Ted Nelson declara en una de sus obras: *"Por hipertexto entiendo escritura no secuencial. La escritura tradicional es secuencial por dos razones. Primero, se deriva del discurso hablado, que es secuencial, y segundo, porque los libros están escritos para leerse de forma secuencial... sin embargo, las estructuras de las ideas no son secuenciales. Están interrelacionadas en múltiples direcciones. Y cuando escribimos siempre tratamos de relacionar cosas de forma no secuencial"* [Nelson, 1974]. En este aspecto, varios son los autores que comparten el argumento de Nelson y a su vez señalan precursores del hipertexto en los textos impresos.

Uno de los mayores referentes, el profesor George P. Landow, describe cómo el hipertexto ha logrado materializar los ideales de intertextualidad perseguidos por pensadores como Barthes, Derrida y Foucault. Según Landow, recursos como las citas y referencias bibliográficas, las notas al margen o pie de página e incluso las tradicionales glosas fallan en su disposición espacial, forzando al lector a redirigir la vista a un nuevo espacio. El hipertexto electrónico, en cambio, facilita la consulta de referencias gracias a la navegación dentro de una misma área de visión [Landow, 1995]. Sin embargo, obras como los diccionarios y enciclopedias han sido inherentemente de lectura no secuencial y sirven de contraejemplo para los críticos literarios.

Entre las definiciones de hipertexto, se entremezcla "hipermedia", un concepto nacido de la inclusión de elementos multimedia al texto electrónico. Aunque los usuarios se han encargado de absorber este concepto dentro del término "hipertexto", es muy común encontrar definiciones compartidas. Nuevamente, fue Nelson quien acuñó "hipermedia" en el mismo artículo en que presentó el hipertexto, donde exponía: *"Las películas y las grabaciones de sonido y video también son hilos lineales, básicamente por razones mecánicas. Pero ahora también pueden ordenarse como sistemas no lineales –en*

retículas, por ejemplo– con el objetivo de editarlas o presentarlas con distintos énfasis. […] El hiperfilm –una película explorable o vari-secuenciada– es sólo una de los posibles hipermedia que demandan nuestra atención"[10] [Nelson, 1965].

En *Hypertext and Hypermedia*, Jakob Nielsen enuncia: *"El hipertexto consiste en piezas de texto o de otro tipo de presentación de la información ligadas de manera no-secuencial. Si el foco de tal sistema descansa en tipos de información no textual, se utiliza el término Hipermedia..."* [Nielsen, 1990].

Por su parte, Landow caracteriza el hipertexto como "un texto compuesto de fragmentos de texto (...) y los nexos electrónicos que lo conectan entre sí" y aclara sobre hipermedia: *"La expresión hipermedia simplemente extiende la noción de texto hipertextual al incluir información visual, sonora, animación y otras formas de información. Puesto que el hipertexto, al poder conectar un pasaje de discurso verbal e imágenes, mapas, diagramas y sonido tan fácilmente como a otro fragmento verbal, expande la noción de texto más allá de lo meramente verbal, no haré la distinción entre hipertexto e hipermedia. Con hipertexto, pues, me referiré a un medio informático que relaciona información tanto verbal como no verbal"* [Landow, 1995].

Las definiciones más simples son quizás las aportadas por el World Wide Web Consortium (el organismo de estandarización de la Web), el cual definió hipertexto como "un texto que no se limita a la linealidad", "un texto con enlaces a otros textos" [W3C, 1995]. En ese mismo documento, hipermedia se define como "hipertexto multimedia; medios distintos al texto que generalmente incluyen gráficos, sonido y video" aunque se menciona la tendencia popular de utilizar ambos términos de manera indistinta.

[10] Texto original: "*Films, sound recordings, and video recordings are also linear strings, basically for mechanical reasons. But these, too, can now be arranged as non-linear systems - for instance, lattices - for editing purposes, or for display with different emphasis. [...] The hyperfilm - a browsable or vari-sequenced movie - is only one of the possible hypermedia that require our attention.*"

Más hacia nuestros días, el exponente de la literatura electrónica Noah Wardrip-Fruin, concluyó: *"Hipertexto es un término acuñado por Ted Nelson para referirse a formas de hipermedia (medios creados por humanos que se ramifican o ejecutan a pedido) que operan en forma textual. Ejemplos de ello son el 'hipertexto discreto' basado en enlaces (del cual la Web es un ejemplo) y el 'stretchtext' basado en niveles de detalle"*[11] [Wardrip-Fruin, 2004].

Por otra parte, la estructura de los hipertextos está claramente definida y ampliamente aceptada. Formalmente, se componen de nodos (recursos vinculados), anclas (fragmentos que dirigen a otros nodos) y enlaces (conexiones entre nodos).

La Tabla 1.1 elaborada por Lamarca Lapuente expone una sintética comparación entre textos tradicionales e hipertextos basada en características de contenido, uso y tecnología.

	TEXTO	**HIPERTEXTO**
Estructura de la información	Secuencial	No secuencial o multisecuencial
Soporte	Papel	Electrónico/Digital
Dispositivo de lectura	Libro	Pantalla
Forma de acceso	Lectura	Navegación
Índice/sumario del contenido	Tabla de contenidos	Mapa de navegación
Morfología del contenido	Texto e imágenes estáticas	Texto, imágenes estáticas y dinámicas, audio, vídeo y procedimientos interactivos
Portabilidad	Fácil de portar y usar	Es necesario disponer de una computadora o un dispositivo especial de lectura
Uso	Puede leerse en cualquier sitio	Para leer se precisa una estación multimedia

Tabla 1.1: Diferencias entre el texto tradicional y el hipertexto. [Lamarca Lapuente, 2007]

[11] Texto original: "*Hypertext is a term coined by Ted Nelson for forms of hypermedia (human-authored media that branch or perform on request) that operate textually. Examples include the link-based 'discrete hypertext' (of which the Web is one example) and the level-of-detail-based 'stretchtext.'*"

En el mismo trabajo, la autora también enuncia aspectos que permiten caracterizar el modelo de hipertexto, aquí resumidos brevemente:

- *Conectividad*

  Cualidad por la cual se establecen conexiones inter e intradocumentales por medio de enlaces, de acuerdo a alguna tipología.

- *Digitalidad*

  Naturaleza digital de los textos electrónicos. Los defensores del hipertexto sostienen que los hiperdocumentos sólo pueden existir en ambientes digitales, ya que permiten acceso veloz y directo a la información referenciada.

- *Multisecuencialidad*

  Hace referencia a la falta de secuencialidad, o según nuevas corrientes la gran variedad de secuencias distintas, en las que puede leerse un hipertexto, en contraposición con la linealidad característica de los textos impresos.

- *Estructura en red*

  Topología descentralizada de los nodos (textos o recursos) que componen un hipertexto, la que obedece al tipo de relaciones que se han representado con los enlaces.

- *Multimedialidad*

  Integración de texto con elementos audiovisuales, como imágenes, animaciones, audio y video.

- *Gradualidad*

  Jerarquización y organización gradual del contenido, en profundidad y extensibilidad.

- *Extensibilidad*

  En realidad, extensibilidad horizontal, que distingue al hipertexto de los medios impresos. Los enlaces o referencias no son necesariamente verticales (hacia arriba o hacia abajo del texto) sino también horizontales (entre hiperdocumentos).

- *Interactividad*

  Grado de interacción y control del usuario sobre el hipertexto, condicionados por la riqueza de funcionalidad y su interfaz gráfica.

- *Usabilidad*

  Diseño, estructura y presentación de un hipertexto, en estrecha relación con la facilidad y efectividad de su uso.

- *Accesibilidad*

  Posibilidad utilización por la mayor cantidad de usuarios posible, independientemente de sus limitaciones físicas, cognitivas, de equipamiento, etc.

- *Reusabilidad*

  Facilidad a la hora de reproducir y modificar los contenidos y sus formatos.

- *Dinamismo*

  Actualización fácil, rápida y continua de los contenidos.

- *Transitoriedad*

  Comprende la inestabilidad, volatilidad y actualización temporal y espacial de los documentos.

- *Apertura*

  Capacidad de modificación y expansión de un hipertexto, en contraste con las limitaciones de los textos impresos, totalmente cerrados y con un principio y fin determinado.

Todas estas perspectivas históricas del concepto de hipertexto e hipermedia convergen en la misma esencia: el hipertexto, como concepto general, comprende textos, imágenes, sonidos, videos y demás recursos multimedia, vinculados entre sí mediante enlaces, que hacen posible la navegación no secuencial de contenido.

1.3. Motivación

Desde sus orígenes, el hipertexto ha revolucionado la lectura, creando una alternativa a los tradicionales textos en papel, caracterizados por su naturaleza estática y lineal. Además de brindar una experiencia mucho más ágil, interactiva y dinámica a sus lectores, el hipertexto presenta un marco ideal para lograr una comprensión de textos más profunda, gracias a la vinculación de segmentos de contenido con recursos explicativos e ilustrativos. De esta manera, es posible dotar a los simples textos de cualidades auto-explicativas, por ejemplo mediante la definición de conceptos, inclusión de información contextual sobre hechos o personajes mencionados, datos descriptivos y elementos multimedia complementarios.

En la actualidad, la World Wide Web pone a disposición de sus usuarios un insuperable conjunto de hipertextos sobre cualquier tema, con vínculos y recursos que ayudan a comprender o profundizar su contenido. El valioso marco contextual capaz de brindar un hipertexto, ha convertido las búsquedas en la Web en una estrategia rápida, práctica y beneficiosa a la hora de aproximarse a temas desconocidos.

Sin embargo, la disponibilidad de hipertextos de interés para los usuarios en el espacio web está condicionada por varios factores. Diversos autores han destacado los siguientes:

a. la falta de contenidos web sobre el tema consultado debido a su gran especificidad, poco conocimiento o difusión,

b. la ausencia de hipertextos con términos o frases específicas de gran interés para el lector,

c. la alta cantidad de contenido engañoso, comercial o irrelevante (denominado "ruido documental") [Piper, 2000],

d. la eventual baja calidad o dudoso contenido de los textos encontrados sobre el tema [Allen, 2002] [Eysenbach, 2002] [Martin-Facklam, 2004] [Bedell, 2004],

e. la disponibilidad de hipertextos que, aún siendo de buena calidad, no generan gran interés en el usuario o simplemente contienen una cantidad de vínculos o recursos insuficientes, poco interesantes, de baja confiabilidad o disponibilidad,

f. la "invisibilidad" de sitios web de alta calidad [Sherman, 2001],

g. la baja disponibilidad de textos en español (sólo un 4,6% de toda la Web) [Accenture, 2006],

h. la incidencia de las herramientas de búsqueda con respecto a los contenidos accedidos, ya sea por las ubicaciones en los ránkings de resultados como por la tendencia de los usuarios a desestimar los resultados que no estén presentes en las primeras páginas de respuestas [iProspect, 2006] [Jansen, 2006],

i. la eventual falta de destreza de los usuarios en la utilización de herramientas de búsqueda, por ejemplo mediante la generación de consultas poco específicas o inexactas [Sisson, 2003] [Rose, 2004], y

j. el hecho de que la existencia de hipertextos esté supeditada a un proceso de creación manual previo por parte de humanos.

Para superar muchas de estas dificultades, los usuarios han confiado principalmente en los motores de búsqueda [Fallows, 2008], aunque éstos no puedan garantizar la satisfacción del usuario ni la calidad de los recursos que brindan. Como alternativa, se

han propuesto tanto consejos prácticos para los usuarios [Tillman, 2003] [Harris, 2007] como modelos computacionales orientados a evaluar la calidad de la información en la Web [Dondio, 2007] [Herrera-Viedma, 2004] [van Gils, 2007]. Sin embargo, los usuarios podrían disponer de una mejor manera de conseguir los hipertextos que más se adecuen a sus necesidades: ser ellos mismos quienes propongan textos base para transformarlos en hipertextos mediante algún mecanismo automático. Este enfoque no implica necesariamente que deban redactar los textos y crear sus hipervínculos sino que su participación podría reducirse simplemente a elegir un texto adecuado y delegar la tarea de generación de hipervínculos a una herramienta automática.

La existencia de una utilidad con estas características permitiría a los usuarios generar sus propios hipertextos a partir de textos de su interés, sin necesidad de buscar ni depender de la existencia de hipertextos en la Web que se ajusten tan rigurosamente a sus necesidades. Esta tarea de conversión automática de texto en hipertexto con la consecuente búsqueda de recursos relevantes y generación de hipervínculos será denominada Enriquecimiento Automático de Textos (EAT) y constituye el objetivo de este trabajo.

1.4. Objetivos

1.4.1. Objetivos Generales

Este trabajo tiene como objetivo proponer, diseñar y desarrollar una arquitectura para el enriquecimiento automático de textos en español, capaz de transformarlos en hipertextos auto-explicativos. Adicionalmente, dentro del ámbito de este trabajo, se incluye el desarrollo de un prototipo funcional con el propósito de validar las formulaciones.

La tarea de enriquecimiento consiste en vincular conceptos y frases dentro de un texto con datos y recursos disponibles en la World Wide Web. Estos elementos textuales, denominados *entidades*, conforman el punto de partida para la búsqueda y selección de recursos y representan los enlaces a la información encontrada.

El principal beneficio del Enriquecimiento Automático de Textos es su capacidad para incorporar recursos a un texto en forma automática. Dicho proceso detecta entidades dentro un texto y las vincula a cuadros de información contextual que aportan datos y recursos complementarios disponibles en la World Wide Web, todo ello sin necesidad de interacción humana.

Los usuarios sólo actúan como disparadores del proceso al aportar el texto que desean enriquecer. Luego, el enriquecedor automático se encarga de la búsqueda, selección y vinculación de los recursos necesarios para generar los hipertextos, utilizando la Web como un repositorio de servicios, información y recursos de manera transparente al usuario.

Una de las características principales de la arquitectura propuesta radica en que se trabaja específicamente con textos en español, lo que plantea desafíos muy interesantes, tanto desde la perspectiva del análisis léxico/semántico de los textos de origen como de disponibilidad de recursos en la Web.

Como se mencionó anteriormente, el proceso de enriquecimiento de textos genera como resultado un hipertexto, cuyas entidades detectadas en el texto original son transformadas en hipervínculos a cuadros con información contextual detallada y recursos web relevantes cuidadosamente seleccionados.

Como ejemplo, considérese el siguiente texto:

Natalia Oreiro junto con Greenpeace convocan a marcha azul por las ballenas

> Buenos Aires, Argentina — La actriz Natalia Oreiro se sumó a la Campaña contra la Caza Comercial de Ballenas de Greenpeace para convocar, a través de un spot que circula por internet, a una gran marcha azul por las ballenas, que se realizará el próximo domingo 27 de mayo a las 14 en el obelisco porteño.
>
> Más información en www.greenpeace.org

La aplicación del enriquecedor propuesto al texto anterior debería identificar las entidades mencionadas y proveer información sobre cada una de ellas según lo definido para cada tipo de entidad. Así, se obtendría la siguiente versión enriquecida (Figura 1.1):

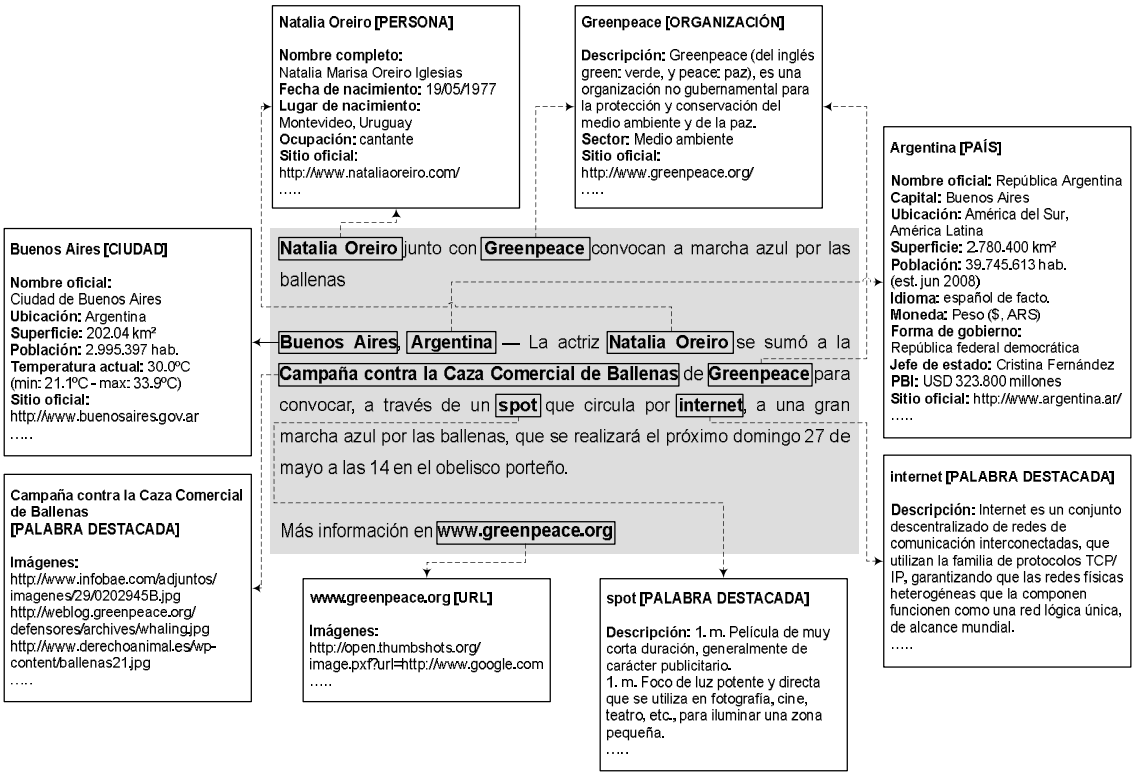

Figura 1.1: Ejemplo de documento enriquecido.

La arquitectura que se presenta en este trabajo está orientada principalmente al enriquecimiento de textos informativos o explicativos por lo que los desarrollos subsiguientes están optimizados para el trabajo con dichos tipos de textos. Sin embargo, esta especificación no limita necesariamente el uso, aplicación ni adaptación de la

arquitectura a otros tipos de texto que pudieran ser enriquecidos de manera satisfactoria.

1.4.2. Objetivos Específicos

Los objetivos generales planteados implican el alcance de los siguientes objetivos específicos:

a. Seleccionar o definir un método de reconocimiento de entidades en español para determinar los conceptos y frases que serán enriquecidas en cada texto.

b. Diseñar un modelo de enriquecimiento para cada tipo de entidad reconocida, determinando los datos y tipos de recursos deseables para cada caso.

c. Investigar, recopilar, analizar y seleccionar un conjunto de fuentes y servicios disponibles en la World Wide Web capaces de proveer los datos y recursos determinados en la etapa anterior.

d. Seleccionar los datos y recursos disponibles en línea que resulten relevantes para cada entidad reconocida en el texto, empleando una estrategia de combinación óptima de datos y recursos provenientes de múltiples fuentes.

e. Diseñar un esquema de presentación apropiada de los resultados, lo que implicará determinar la cantidad adecuada de entidades a enriquecer, analizar las formas más apropiadas de presentación del enriquecimiento, facilitar la navegación del texto y los recursos integrados, entre otros.

f. Implementar la arquitectura de enriquecimiento automático de textos como un prototipo de aplicación.

1.4.3. Destinatarios y Aplicaciones

La arquitectura propuesta en este trabajo está orientada principalmente al enriquecimiento de textos informativos o explicativos en español, tales como noticias, artículos periodísticos, descripciones de personajes, biografías, textos históricos, reseñas, monografías, relatos, etcétera, aunque dicha orientación no limita la aplicación del enriquecimiento propuesto a otros tipos de textos donde también pueda resultar beneficioso.

La utilidad brindada por el enriquecimiento de textos puede ser aprovechada por distintos tipos de usuarios en contextos donde la comprensión integral y profunda de los contenidos sea lo primordial.

Entre los usuarios y situaciones típicas de aplicación pueden mencionarse:

- estudiantes que deseen enriquecer los textos escolares para facilitar la compresión del contenido y obtener información sobre conceptos desconocidos del texto,
- lectores de artículos de actualidad que requieran información detallada y actualizada sobre personajes o acontecimientos mencionados en el texto,
- usuarios que deseen evitar múltiples búsquedas de definiciones o recursos relacionados con conceptos importantes de un texto,
- lectores que deseen complementar sus lecturas con recursos multimedia,
- usuarios en busca de una herramienta informativa contextual.

Aunque la arquitectura propuesta es implementada en un prototipo funcional orientado directamente a usuarios finales, también podría implementarse como complemento a otros procesos o aplicaciones existentes, donde podría constituir una capa de software

encargada de realizar enriquecimientos. Ejemplos de ello podrían ser el enriquecimiento de artículos de periódicos disponibles en la Web o incluso de mensajes de correo electrónico.

1.5. Áreas Involucradas

Para la realización de este trabajo, es necesario aplicar técnicas y recursos provenientes de diversas áreas de de las Ciencias de la Computación. Por tal motivo, se brinda a continuación una breve descripción de cada una de las áreas involucradas, que serán descriptas con mayor detalle en el resto del trabajo.

1.5.1. Reconocimiento de Entidades

El primer paso dentro del enriquecimiento de textos consiste en determinar qué se va a enriquecer, es decir, identificar los conceptos, frases o *entidades* mencionadas en el texto que se enriquecerán.

Se denominan *entidades* a los elementos del texto que corresponden a alguna categoría predefinida de objetos, generalmente nombres propios, como pueden ser una ciudad, un país, una organización, una persona, una marca registrada, etcétera.

El área en la cual se investigan los modelos y técnicas de reconocimiento de entidades dentro de un texto se conoce como Reconocimiento de Entidades con Nombre (*Named Entity Recognition* o NER) [Chinchor, 1995] y es una rama de Extracción de Información.

En este trabajo se analizan algunas de las técnicas más populares de esta área, se describe su rol dentro del proceso de enriquecimiento y se implementa un híbrido de técnicas que permita resolver este primer paso en la etapa de desarrollo.

1.5.2. Enriquecimiento de Entidades

Una vez reconocidos los elementos del texto sobre los que se trabajará, debe procederse a su enriquecimiento. Como se explica más adelante, no existe un completo consenso entre los investigadores acerca de qué implica el enriquecimiento de entidades ni cómo debe llevarse a cabo pues, de acuerdo a lo investigado, no existe un área de investigación formal dedicada a esta tarea. De todos modos, existen enfoques o técnicas propuestas por algunos autores que apuntan a objetivos similares por lo que se parte de esas contribuciones para intentar definir formalmente en qué consistiría un Enriquecimiento de Entidades.

Básicamente, la tarea implica la búsqueda, recopilación, extracción e integración de recursos e información relativos a una entidad dentro de uno o más repositorios de información. Para este trabajo, las entidades enriquecidas son aquellas reconocidas en la etapa anterior mientras que el repositorio de servicios e información consultado es conformado por la World Wide Web.

1.5.3. Presentación de Resultados

La fase final de este trabajo reside en la investigación y adaptación de un marco de presentación de resultados adecuado para la información de enriquecimiento obtenida en las etapas anteriores.

De acuerdo a lo planteado oportunamente, el resultado final del proceso de enriquecimiento propuesto en este trabajo consiste en una versión hipertextual del texto inicial. Las entidades marcadas en el nuevo hipertexto son vinculadas a cuadros con información contextual y recursos relacionados que aportan mayor riqueza informativa al contenido. Para lograr una adecuada transformación del texto en hipertexto deben analizarse los trabajos provenientes de las áreas de Visualización de Información (o Infovis) [Spence, 2000] [Fluit, 2002] e Interacción Hombre-Máquina (Human Computer Interaction, HCI) [Myers, 1998]. A través de dichas áreas se intenta determinar cánones para una óptima presentación del hipertexto generado, las entidades reconocidas, su información contextual y los recursos adicionales recopilados.

La aplicación de los conocimientos de estas áreas a la arquitectura propuesta determina el producto final que recibirá el usuario como resultado del enriquecimiento de texto, respetando los principios de usabilidad y accesibilidad, entre otros.