UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERIA INDUSTRIAL
DEPARTAMENTO DE CIENCIAS DE LA COMPUTACIÓN

DISEÑO E IMPLEMENTACIÓN DE UN FRAMEWORK PARA LLEVAR SITIOS WEB SOCIALES A LA WEB SEMÁNTICA Y QUE FACILITE EL ANÁLISIS DE REDES SOCIALES.

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL EN COMPUTACIÓN E INGENIERO CIVIL INDUSTRIAL

FRANCISCO JAVIER BUSTOS CARVAJAL

PROFESOR GUÍA:
SEBASTIÁN RÍOS PÉREZ

MIEMBROS DE LA COMISIÓN:
CLAUDIO GUTIÉRREZ GALLARDO
SERGIO OCHOA DELORENZI
GASTÓN L'HUILLIER CHAPARRO

SANTIAGO DE CHILE
JULIO 2011.

RESUMEN DE LA MEMORIA
PARA OPTAR AL TÍTULO DE
INGENIERO CIVIL EN COMPUTACIÓN E
INGENIERO CIVIL INDUSTRIAL
POR: FRANCISCO BUSTOS CARVAJAL
FECHA: 20/06/2011
PROF. GUIA: SR. SEBASTIÁN RÍOS.

DISEÑO E IMPLEMENTACIÓN DE UN FRAMEWORK PARA LLEVAR SITIOS WEB SOCIALES A LA WEB SEMÁNTICA Y QUE FACILITE EL ANÁLISIS DE REDES SOCIALES

Unos de los sitios web sociales con mayor auge en el último tiempo ha sido Facebook, el cual actualmente tiene más de 500 millones de usuarios activos[1]. Sin embargo, a pesar del fuerte surgimiento de los sitios web sociales, estos conviven separados entre sí. No existe integración o reutilización de la información entre diferentes sitios, ni tampoco existen incentivos para hacerlo.

Según Kinsella et al. [16], la Web Semántica podría ayudar a crear sitios web sociales interconectados e interoperables. Sin embargo, el primer paso es llevar los actuales sitios web a la Web Semántica, para luego crear vínculos entre éstos. Algunos de los proyectos relacionados con transformar sitios web sociales en sitios web sociales semánticos son: SIOC (Semantically-Interlinked Online Communities), FOAF (friend of a friend), entre otros.

La Web Semántica, también podría facilitar y enriquecer el análisis de redes sociales sobre sitios web sociales, mediante el estudio de múltiples sitios simultáneamente y de manera transparente.

Para llevar un sitio web social a la Web Semántica, se requiere conocer detalles de la arquitectura y del modelo de datos del sitio a migrar, además de la especificación de la Web Semántica. De esta manera, surge la necesidad de crear elementos de software reutilizables que permitan servir de apoyo en esta labor.

Este trabajo tuvo por objetivo diseñar e implementar un framework que permita llevar foros web a la Web Semántica y que sirva de apoyo para el estudio de éstos en base al análisis de redes sociales. El framework fue validado mediante el desarrollo de una aplicación que permitió llevar un foro web real a la Web Semántica. Luego, utilizando una herramienta para el análisis de redes sociales, se calcularon diversos indicadores para medir el comportamiento de sus miembros.

Finalmente, como una forma de contribuir a la comunidad global para fomentar el uso de la Web Semántica, se publicó el framework en el sitio SourceForge[2], bajo la licencia GNU General Public Licence.

[1] http://www.facebook.com/press/info.php?statistics [Fecha último acceso: marzo de 2011]
[2] http://sourceforge.net/projects/siocextended/ [Fecha último acceso: marzo de 2011]

Índice general

Índice de figuras

Índice de cuadros

Lista de Anexos

Capítulo 1

Introducción

Hoy en día es natural leer el diario, buscar empleo, comprar, compartir información de interés con otros y hacer amigos a través de la Web. Sin embargo, esto no siempre fue posible. En un comienzo, los sitios web no permitían que sus usuarios (o mejor dicho "lectores") fuesen parte activa de la generación del contenido. Sólo un pequeño grupo de personas tenían el privilegio de generar contenido.

Sin embargo, los usuarios de la Web no sólo querían ser lectores, sino también interactuar con las ideas y pensamientos de otros [22]. Este fenómeno social generó un cambio radical en la Web, que está marcado por la aparición de los sitios web sociales, como por ejemplo: Youtube, Blogger, Twitter, Facebook, Flikr, Wikepedia, entre otros. Los sitios web sociales se caracterizan porque ofrecen diversas aplicaciones para que sus miembros generen contenido e interactúen entre ellos.

Un sitio web social ícono es Facebook[1], el que actualmente tiene más de 500 millones de usuarios activos, y que alrededor del 70 % de esos usuarios no son de Estados Unidos. Un usuario promedio de Facebook tiene 130 amigos y crea 90 piezas de contenido por mes. Estos datos dan una vislumbre del poder de los sitios web sociales para relacionar personas.

A pesar de su explosivo crecimiento, los sitios web sociales coexisten aislados entre sí. Hoy en día, no es posible la integración y reutilización de la información entre diferentes sitios. Por ejemplo: una persona para ser miembro en diferente sitios, debe registrarse en cada uno de ellos; éste no puede reutilizar su información de perfil de un sitio en otro. Esto pone en evidencia que los sitios web sociales no fueron construidos para colaborar entre ellos.

[1] http://www.facebook.com/press/info.php?statistics [Fecha último acceso: marzo de 2011]

Según Kinsella et al. [16], los problemas de la Web Social son de origen tecnológicos. La Web fue creada en base la publicación de documentos HTML (HyperText Markup Language) y no en base a los datos. La información expuesta en documentos HTML sólo puede ser interpretada por un ser humano y no por máquinas. Los robots no pueden tomar decisiones a partir de la información publicada en HTML, ya que no la pueden interpretar.

La Web Semántica surgió como una respuesta a los problemas tecnológicos de la Web. El término Web Semántica fue introducido por Tim Berners-Lee, principal inventor de la Web, en el año 2001 [3]. Berners-Lee Definió la Web Semántica como la Web en que sus datos pueden ser procesados no sólo por humanos, sino también por máquinas. En ese sentido, la Web Semántica provee las bases tecnológicas para la generación de sitios web sociales interconectados e interoperables.

La Web Semántica no sólo permitiría potenciar los sitios web sociales, sino también abre la posibilidad de enriquecer y facilitar el estudio de éstos. Una de las mayores dificultades para el análisis de los sitios web sociales, es que no es posible reutilizar algoritmos y/o aplicaciones desarrolladas para un sitio particular en otros sitios, ya que cada sitio web maneja su propia estructura de almacenamiento de los datos. La Web Semántica, inherentemente, ofrece una representación estándar para cada dominio específico, lo que permitiría un mayor nivel de automatización y reutilización de los algoritmos y/o aplicaciones desarrolladas para el análisis de los sitios web sociales.

Además, bajo el supuesto de contar con sitios web sociales interoperables e interconectados, sería posible realizar estudios sobre la información agregada de diferentes sitios web de manera transparente. Por ejemplo, se podría aplicar análisis de redes sociales (SNA) sobre múltiples sitios.

La Figura 1.1 muestra un resumen de los beneficios de aplicar la Web Semántica a la Web Social. A pesar que, la era de la Web Semántica no es un realidad todavía, existen diversos trabajos alrededor del mundo que pretenden cimentar el camino entre la Web Social y la Web Semántica, por ejemplo: el proyecto Linking Open Data, cuyo objetivo es crear un respositorio de datos abiertos y ligados entre sí; FOAF[2] (friend of a friend), cuyo objetivo es la representación semántica de los datos de las personas y sus relaciones; SIOC[3] (Semantically-Interlinked Online Communities), que tiene por objetivo la integración de comunidades virtuales mediante tecnologías de la Web Semántica, entre otras.

Este trabajo también representa un aporte para la migración de la Web Social a la

[2]http://www.foaf-project.org/ [Fecha último acceso: marzo de 2011]
[3]http://sioc-project.org/ [Fecha último acceso: marzo de 2011]

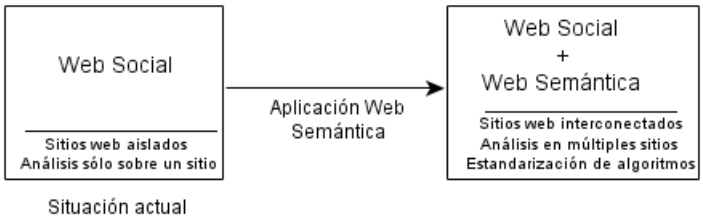


Figura 1.1: Beneficios de la Web Semántica

Web Semántica, pero limitado a los foros web. Primero se diseño e implementó un framework que permite llevar foros web a la Web Semántica en base a la ontología de SIOC. Y Luego, se extendió el framework para que fuese posible el análisis de redes sociales sobre la estructura semántica generada por éste.

Para validar el trabajo realizado, se utilizó el framework para llevar un foro web real a la Web Semántica, y luego se obtuvieron algunos indicadores aplicando análisis de redes sociales.

1.1. Justificación

El hecho de contribuir a la Web Semántica fue una primera justificación para el desarrollo de este trabajo, no obstante, a continuación se detallan otros argumentos considerados.

¿Por qué la Web Semántica? Los sitios web sociales, a pesar de su explosivo crecimiento, están aislados y desconectados entre sí. Esto se debe principalmente a la que la Web se creó en base a la publicación de información mediante documentos HTML y no en base a la publicación de información interpretable por una máquina. El estándar propuesto por la Web Semántica permite crear sitios web sociales interconectados e interoperables. Además, por el hecho de contar con una representación estándar de la información, sería posible estandarizar algoritmos y/o aplicaciones que son base para el estudio y mejoras de los sitios web sociales.

¿Por qué se desarrolló un framework? Migrar sitios web a la Web Semántica no es una tarea sencilla. Se requiere mucho esfuerzo para comprender los fundamentos de la Web Semántica y las tecnologías asociadas. El framework permite empaquetar todo ese conocimiento, para que el trabajo de los desarrolladores sea más simple y estándar, y así facilitar la migración de foros web a la Web Semántica.

¿Por qué se utilizó SIOC? Porque SIOC desarrolló una ontología o modelo para representar e intercontecar sitios web sociales mediante el uso de la Web Semántica. Por otro lado, SIOC fue aceptado como parte de la W3C[4] (World Wide Web Consortium), organización mundial encargada de generar los estándares de la Web, en el año 2007[5], por lo tanto éste cumple con los requisitos de la Web Semántica.

Si SIOC es tan bueno, ¿Acaso no existe un framework para llevar sitios web sociales a la Web Semántica? La respuesta es no. SIOC ha desarrollado varios exportadores para foros, WebLog y CMS (por ejemplo: WordPress SIOC Exporter, DotcLear SIOC Exporter, Drupal SIOC Exporter, PHPBB2 SIOC Exporter, etc.), sin embargo, son aplicaciones individuales que van embebidas dentro de cada sitio web social y por lo tanto, desarrolladas en el mismo lenguaje de programación del sitio web.

En este trabajo se desarrolló un framework que define un marco de trabajo estándar, en que todo el conocimiento se encuentra embebido en el mismo framework y no lo en los exportadores. Además, se implementaron componentes de software para que el framework genere mecanismos para aplicar análisis de redes sociales sobre la estructura semántica generada por éste.

1.2. Objetivos

Objetivo General

Diseñar e implementar un framework que permita llevar foros web a la Web Semántica y que entregue mecanismos para aplicar análisis de redes sociales sobre las estructuras semánticas generadas por éste. Y mediante los componentes provistos por el framework, desarrollar una aplicación prototipo.

Objetivos Específicos

1. Realizar un levantamiento de la situación actual sobre la Web Social y la Web Semántica.

2. Realizar un estudio referente al análisis de redes sociales.

[4]http://www.w3.org/ [Fecha último acceso: marzo de 2011]
[5]http://www.w3.org/Submission/2007/02/ [Fecha último acceso: marzo de 2011]

3. Realizar un estudio acerca de los tipos de frameworks existentes, sus principales características y las mejores prácticas de desarrollo.

4. Diseñar e implementar un conjunto de herramientas de software que permitan llevar los foros web a la Web Semántica.

5. Diseñar e implementar un conjunto de herramientas de software que entreguen los mecanismos necesarios para aplicar análisis de redes sociales sobre la estructura semántica generada por el framework.

6. Validar el trabajo realizado mediante la construcción de una aplicación que permita, en base a los componentes de software desarrollados, llevar un foro web social real a la Web Semántica y aplicar análisis de redes sociales sobre éste.

7. Contribuir a la comunidad global para fomentar el uso de la Web Semántica.

1.3. Resultados Esperados

A continuación se detallan los resultados esperados, a partir de cada objetivo específico.

Mediante el objetivo específico: "Realizar un levantamiento de la situación actual sobre la Web Social y la Web Semántica**",** se intenta obtener:

1. Una sección del capítulo de antecedentes, en la que se detalle los grandes movimientos que ha tenido la Web y los cambios que se esperan para el futuro.

2. Una sección del capítulo de antecedentes, en la que se detalle como la Web Semántica puede ayudar a resolver los problemas de interoperabilidad de los sitios web sociales y enriquecer el estudio de estos, en base al análisis de redes sociales.

Mediante el objetivo específico: "Realizar un estudio referente al análisis de redes sociales", se intenta obtener:

1. Una sección del capítulo de antecedentes, en la que se detallen los principales indicadores utilizados para el análisis de redes sociales.

2. Una sección del capítulo de antecedentes, en la que se explique el algoritmo Hits y

Pagerank.

3. Una sección del capítulo de antecendetes, en la que se muestre como configurar una red social (para aplicar SNA) a partir de un foro.

Mediante el objetivo específico: "Realizar un estudio acerca de los tipos de frameworks existentes, sus principales características y las mejores prácticas de desarrollo**",** se intenta obtener:

1. Una sección del capítulo de antecedentes, en la que se detallen los tipos de frameworks existentes de acuerdo a la forma de extensión (frameworks de caja blanca y caja negra).

2. Una sección del capítulo de antecedentes, en la que se explique los conceptos de hotspots y frozenspots.

3. Una sección en que se mencionen buenas prácticas con respecto al proceso de desarrollo de frameworks.

Mediante el objetivo específico: "Diseñar e implementar un conjunto de herramientas de software que permitan llevar los foros web a la Web Semántica"**,** se intenta obtener:

1. Una API (Application Programming Interface) que permita transformar los datos de un foro a una representación semántica.

2. Un exportador de PHPBB2[6] , que en base a la API, permita llevar los foros en PHPBB2 a la Web Semántica.

Mediante el objetivo específico: **"**Diseñar e implementar un conjunto de herramientas de software que entreguen los mecanismos necesarios para aplicar análisis de redes sociales sobre la estructura semántica generada por el framework**",** se intenta obtener: Un exportador que permita generar las redes sociales o grafos SNA (Social Network Analysis) a partir de la representación semántica para los foros web provista por el framework. Luego, mediante alguna herramienta existente para realizar análisis de redes sociales, como por ejemplo: Gephi[7], se pueden importar las redes de los foros y aplicar SNA sobre éstos.

Mediante el objetivo específico: "Validar el trabajo realizado mediante la construcción de una aplicación que permita, en base a los componentes de software desarrollados, llevar un foro web social real a la Web Semántica y aplicar análisis de redes sociales sobre

[6]http://www.phpbb.com/ [Fecha último acceso: marzo de 2011]
[7]http://gephi.org/ [Fecha último acceso: junio de 2011]

éste", se intenta obtener una aplicación basada en el framework que permita representar semánticamente el foro web Plexilandia[8]. Dicha comunidad está enfocada en las personas que se interesen por construir todo tipo de equipos de sonido, efectos para instrumentos y guitarras. Cabe mencionar que el foro está implementado en PHPBB2.

Luego, se debe aplicar análisis de redes sociales sobre la representación semántica de Plexilandia para obtener el valor de algunos indicadores básicos, como por ejemplo: la importancia de los usuarios, densidad, centralidad, entre otros.

Para: "Contribuir a la comunidad global para fomentar el uso de la Web Semántica", se debe publicar el proyecto en un repositorio de aplicaciones de código abierto, como por ejemplo: SourceForge[9]. Así cualquier desarrollador o investigador puede tener accesso al framework y a su código fuente gratuitamente.

1.4. Alcances

Este trabajo contempla el desarrollo de una API, un exportador para PHPBB2, y un módulo para generar las redes sociales a partir de los datos semánticos. Si bien, existen otros motores de foros de código abierto (Wordpress, PHP Nuke, PHPWCMS, Drupal), éste trabajo se limita sólo a PHPBB2, lo que a su vez permitirá (en el futuro) implementar exportadores para otras plataformas.

Los alcances con respecto a la implementación de la API son: el desarrollo de una API en régimen perezoso y ansioso. En régimen perezoso, los datos son generados a medida que éstos son requeridos. En régimen ansioso, todos los datos son generados de una sola vez.

Los datos de los foros Web que deben representarse semánticamente son: los atributos de los usuarios, como por ejemplo: nombre de usuario e identificador; la información de los mensajes, como por ejemplo: la fecha creación, contenido, respuestas, autor, entre otros; la información global de los foros y/o categorías; y la información general del sitio, como por ejemplo: la página web donde está alojado y las categorías en que se divide.

Para justificar la utilidad del framework se deben generar los grafos para aplicar análisis de redes sociales, a partir de la estructura semántica. No se considera calcular indica-

[8] http://www.plexilandia.cl/foro/index.php [Fecha último acceso: marzo de 2011]
[9] http://sourceforge.net/ [Fecha último acceso: marzo de 2011]

dores de SNA o implementar algún algoritmo de minería de datos. Sólo se debe exportar los grafos a un tipo de archivo, que pueda ser reconocido por alguna herramienta para el análisis de redes sociales, como por ejemplo: los archivos PAJEK.

1.5. Metodología

Para el desarrollo del capítulo de los antecedentes se contempla la lectura de literatura relacionada a la Web Semántica, papers, revistas, artículos científicos y la especificación de la Web Semántica de la W3C[10]

Para el desarrollo del framework se usará la metodología de desarrollo de software asistida por pruebas (Test Driven Development). Esta metodología se basa en los siguientes principios[11]:

1. Primero se escribe una prueba y se verifica que falle

2. Luego se desarrolla el mínimo de código para pasar la prueba

3. Y finalmente se refactoriza el código escrito.

Para el desarrollo de esta metodología es necesario representar cada requerimiento como un conjunto de pruebas. El objetivo de esta metodología es evitar el código innecesario y crear un código limpio.

Se escogió esta metodología porque el desarrollo de frameworks es un caso particular del desarrollo de software. Un framework es una arquitectura incompleta, por lo que es necesario ir verificando el correcto funcionamiento de cada funcionalidad al momento de su desarrollo. Si algún módulo contiene errores, es muy díficil encontrarlos cuando se desarrollan aplicaciones en base al framework.

[10] http://www.w3.org/standards/semanticweb/ [Fecha último acceso: marzo de 2011]
[11] http://en.wikipedia.org/wiki/Test-driven_development [Fecha último acceso: marzo de 2011]

Capítulo 2

Antecedentes

En este capítulo se entregan todos los conceptos involucrados con el trabajo desarrollado. En la sección 2.1 se realiza un estudio de la situación actual de la Web y su evolución desde su creación hasta la Web Semántica. En la sección 2.2 se exponen los principales conceptos del análisis de redes sociales y como se aplica a los sitios web sociales. Y finalmente, en la sección 2.3 se recopilan los antecedentes o estado del arte con respecto al desarrollo frameworks, los tipos de frameworks y las mejores prácticas de desarrollo.

2.1. La Web

Hoy en día es natural enviar y recibir mensajes, buscar empleo, comprar, compartir información de interés con otros y hacer amigos a través de la Web. Sin embargo esto no siempre fue así. En un principio la Web los usuarios de la Web no tenían los privilegios de generar contenido, sólo eran "lectores".

No obstante, con el tiempo la Web fue evolucionando. Alrededor del año 2000, comenzaron a surgir sitios web que permitían a sus usuarios ser parte activa en la generación de contenido. Este tipo de sitios son llamados sitios web sociales. Algunos ejemplos de sitios web sociales, que han tenido gran éxito en el último tiempo, son: Blogger, Twitter, Facebook, Flikr, Wikepedia, Youtube, entre otros. A este nuevo movimiento de la Web se le conoce como la Web 2.0 o la Web Social.

Sin embargo, una limitación de los sitios web sociales es que están aislados entre sí,

como islas en el mar. Luego, no es sencillo integrar el conocimiento de dos sitios web sociales automáticamente. Por otro lado, en la Web actual, un usuario no puede reutilizar la información de su perfil en diferentes sitios, sino debe registrarse en cada sitio. De aquí se desprende que los sitios web sociales no fueron desarrollados para compartir información e interoperar entre ellos. Además, según Pollock [22], tampoco existen incentivos para hacerlo. Estos problemas se deben básicamente a que la Web fue construida para compartir documentos y no datos.

La Web Semántica pretende resolver estas dificultades mediante una representación semántica de los datos, que inclusive las máquinas puedan interpretar. Tim Berners-Lee, principal inventor de la Web, el año 2001 en [3] introdujo el término de Web Semántica y la definió como "una extensión de la web, en donde la información tiene un significado bien definido, permitiendo la cooperación entre máquinas y personas". Después de varios años, la W3C[1](World Wide Web Consortium), organización mundial encargada de generar los estándares de la Web, publicó los primeros estándares de la Web Semántica.

En los siguientes apartados se registra más en detalle los conceptos de Web Social, los fundamentos de la Web Semántica y el resultado de aplicar la Web Semántica a la Web Social.

2.1.1. La Web Social

Desde su creación, la Web no solo ha facilitado la comunicación entre los computadores, sino también entre las personas. La Web ha permitido romper las barreras geográficas, étnicas, culturales y sociales [6].

Los primeros sitios web (o sitios web 1.0) se caracterizaban porque sus usuarios sólo podían leer el contenido y no tenían el privilegio de ser parte activa de su generación. Por lo tanto el contenido de los sitios web 1.0 no variaba mucho. Sin embargo, la necesidad de los usuarios por ser parte activa en la generación de contenido [15], generó una nueva tendencia en la Web, la que se conoce como la Web Social o la Web 2.0. La Web Social está marcada por la aparición sitios como Youtube, Blogger, Twitter, Facebook, Flikr, Wikepedia, entre otros.

La colaboración y participación de los miembros de los sitios web sociales han permitido, como por ejemplo, que Wikepedia sea la enciclopedia más grande del mundo [12].

[1]http://www.w3.org/ [Fecha último acceso: marzo de 2011]

Otro caso emblemático es Facebook que, actualmente tiene más de 500 millones de usuarios activos[2] y que alrededor del 70 % de esos usuarios no son de Estados Unidos.

El término Web 2.0 fue definido por Tim O'Reilly el año 2005. La Web 2.0 no se refiere a una mejora tecnológica de la Web 1.0, sino que es producto de un fenómeno social [4]. Dentro de los sitios web sociales se encuentran los blogs, wikis, foros, redes sociales online, etc. La Figura 2.1 muestra un esquema simple de la Web Social, en donde Post y Comment representan todos los medios de comunicación sociales (videos, foros, fotos, etc.)

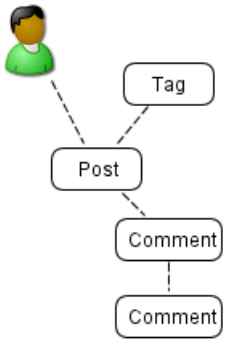


Figura 2.1: La Web Social en términos simples.

Producto de la colaboración y participación entre los miembros de un sitio web social, es posible encontrar comunidades dentro de los mismos sitios, las que se conocen como comunidades virtuales. Una comunidad virtual es un conjunto de personas que comparten intereses, gustos y/o necesidades comunes [18, 8, 30, 31, 27]. Los tipos de comunidades virtuales más conocidos son las comunidades virtuales de práctica, propósito e interés.

- *Comunidades Virtuales de Práctica*: Las comunidades virtuales de práctica son comunidades en que sus miembros comparten sus conocimientos de un tema específico. Las comunidades de prácticas son miradas como fábricas de conocimiento [21, 11, 1, 23]. Algunos ejemplos de comunidades de práctica son: foros relacionados con la ingeniería en sonido, foros relacionados enfocados a desarrolladores de páginas web, entre otros.

- *Comunidades Virtuales de Propósito*: Las comunidades virtuales de propósito son aquellas en que sus miembros comparten objetivos similares, en general son objetivos de corto plazo. Los miembros permanecen en la comunidad hasta que cumplen sus objetivos [32]. Por ejemplo: comunidades para buscar trabajo, comunidades para compradores de viviendas o automóviles.

- *Comunidades Virtuales de Interés*: Las comunidades de interés son aquellas en que sus

[2]http://www.facebook.com/press/info.php?statistics [Fecha último acceso: marzo de 2011]

miembros comparten los mismos intereses. Por ejemplo: el fan club de un grupo musical [29], un grupo religioso, un grupo de alguna barra de futbol, entre otros.

2.1.2. La Web Semántica

Desde su creación, La Web se ha basado en la publicación de información en documentos HTML (HyperText Markup Language), estándar que permite mostrar la información de manera ordena y sencilla. Sin embargo, los documentos HTML son sólo interpretables por un ser humano y no por una máquina.

La Web Semántica propone la publicación de información mediante una tecnología que el hombre y las máquinas puedan interpretar. En ese sentido la Web Semántica propone una mejora tecnológica a la Web. Después de que Tim Bernes-Lee introdujera el concepto de Web Semántica en el año 2001 [3], pasaron varios años para que la W3C publicara la primera especificación de la Web Semántica. A pesar que se han realizado diversos trabajos en base a la Web Semántica, la migración de la Web actual a la Web Semántica se encuentra en la parte más empinada de la curva.

La actual especificación de la Web Semántica se basa en las siguientes tecnologías: RDF[3](Resoruce Description Framework), SPARQL[4](Query Language for RDF), OWL[5](Web Ontology Language), SKOS[6](Simple Knowledge Organization System), entre otras. A continuación se hará un pequeño resumen de algunas, sin embargo, para conocerlas más en detalle, se recomienda revisar la especificación publicada por la W3C.

2.1.2.1. Resource Description Framework (RDF)

RDF es un framework que permite representar la información en la Web de manera que se interpretable por una máquina y un ser humano. RDF se basa en que cada recurso se representa mediante una URI (Uniform Resource Indentifier) y que cada expresión se representa mediante tripletas (grafos RDF). Una tripleta es una forma de relacionar recursos entre sí y de asignar propiedades a los recursos. Una tripleta se compone de tres partes: sujeto, predicado y objeto (ver Figura 2.2). El predicado permite representar una relación

[3]http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/#section-Introduction [Fecha último acceso: marzo de 2011]
[4]http://www.w3.org/TR/2008/REC-rdf-sparql-query-20080115/ [Fecha último acceso: marzo de 2011]
[5]http://www.w3.org/TR/2009/REC-owl2-syntax-20091027/ [Fecha último acceso: marzo de 2011]
[6]http://www.w3.org/TR/2009/REC-skos-reference-20090818/ [Fecha último acceso: marzo de 2011]

entre el sujeto y el objeto, o bien una propiedad del sujeto.

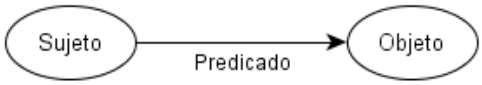


Figura 2.2: Esquema de una tripleta

En la Figura 2.3 se muestra un ejemplo de un grafo RDF con sólo una tripleta, que relaciona al recurso http://www.ejemplo.cl/post1 con una propiedad (un título) igual a "Hola, soy nuevo en el foro". En la figura, se dibujó al sujeto con una elipse y al objeto con un rectángulo, para diferenciar un recurso de un literal.

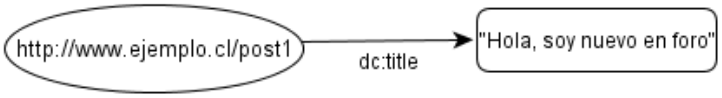


Figura 2.3: Representación de un post mediante un grafo RDF

Un recurso es cualquier cosa representada por una URI, mientras que un literal es sólo una cadena de caracteres. El sujeto puede ser una URI o un nodo en blanco, el predicado sólo puede ser una URI y el objeto puede ser una URI, un nodo en blanco o un literal.

La especificación de RDF muestra como representar un grafo RDF en XML, sin embargo, existen otras formas de representar RDF, como por ejemplo: N-Triples y N3. A continuación se muestra un ejemplo de RDF en formato RDF/XML:

```
<?xml version="1.0"?>
<rdf:RDF
xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:dc="http://purl.org/dc/elements/1.1/">
<rdf:Description rdf:about="http://www.ejemplo.cl/post1">
   <dc:title>"Hola, soy nuevo en el foro"</dc:title>
</rdf:Description>
</rdf:RDF>
```

El siguiente ejemplo corresponde a un RDF en formato N3:

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix dcterms: <http://purl.org/dc/terms/> .
<http://www.ejemplo.cl/post1> dc:title "Hola, soy nuevo en el
   foro"@es;
```

La etiqueta @prefix permite simplificar y factorizar el uso de propiedades o recuros en un archivo RDF. Por ejemplo, dc:title es equivalente a escribir http://purl.org/dc/terms/title. En N3 los recursos se escriben entre corchetes (<aquí va una URI>) y los literales se escriben entre comillas ("esto es un literal").

2.1.2.2. SPARQL

SPARQL es el lenguaje de consulta, recomendado por la W3C, para RDF. La ventaja que tiene SPARQL sobre SQL, es que no está limitado para realizar consultas sobre un sitio. Dado que RDF está fundamentado en el uso de tripletas, SPARQL también.

A continuación se muestra como extraer el título del post1 del ejemplo de la Figura 2.3:

```
PREFIX dcterms: <http://purl.org/dc/terms/>
SELECT ?title
WHERE { http://www.ejemplo.cl/post1> dc:title ?title }
```

La Sintaxis de SPARQL es parecida a la de SQL, sin embargo existen algunas diferencias. La condición dentro de la cláusula WHERE se expresa mediante una tripleta. El sujeto de la tripleta condición es una variable llamada "title". Las variables no tienen valores definidos, por lo tanto sirven para ser retornadas o para hacer uniones entre tripletas. Finalmente el resultado de la consulta es: title= "Hola, soy nuevo en el foro".

2.1.3. La Web Social Semántica

En la Figura 2.4 se muestra la evolución de la Web en sus dos variables: Social y Tecnológico. En un principio la Web se basaba en contendio estático, Después, producto de un fenómeno social, surgió la Web Social. Luego, surgió el concepto de la Web Semántica. Sin embargo, la historia no termina ahí, existe otra dimensión de la Web, conocida como Web Social Semántica, que se basa en la migración de la Web Social a la Web Semántica.

¿Será la Web Social Semántica la Web 4.0 o es parte de la Web 3.0?. No es sencilla la respuesta. Tal vez el problema radica en versionar la Web, ya que su evolución no ha sido sobre el mismo eje. Sin embargo, esta discusión no es parte de este trabajo, por lo que queda propuesta para futuras investigaciones.

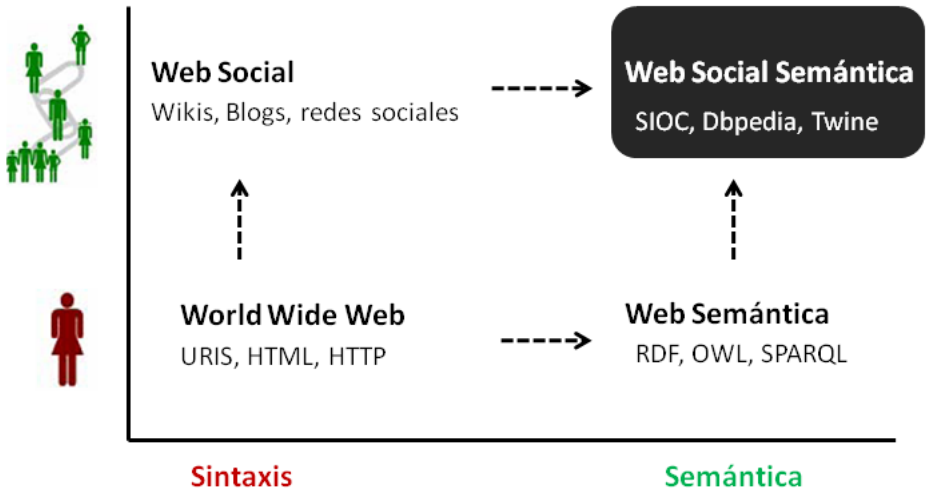


Figura 2.4: Web Social Semántica

2.1.3.1. Web Social Semántica: Más que la suma de sus partes

Según Kinsella et al. [16], la Web Social Semántica es más que la suma entre sus partes, en el sentido en que ambas (La Web Social y la Web Semántica) se ven favorecidas por su unión.

La Semántica puede contribuir a la Web Social para: (1) crear sitios interoperables entre sí. (2) Dar mayor portabilidad a los datos de las personas, lo que se traduce en que las personas sean dueñas de sus datos. (3) Lograr reutilizar el contenido a través de distintos sitios, etc.

La Web Social puede contribuir a la Web Semántica mediante el rico contenido generado por los usuarios. Es común, en los sitios web sociales, hacer etiquetas para describir el contenido de una foto, video, post, etc. Dichas etiquetas pueden facilitar la tarea de migrar la Web a la Web semántica, mediante la extracción de ontologías utilizando técnicas de minería de datos [10].

2.1.3.2. La cadena alimenticia en la Web Social Semántica

En [6] se ilustra el flujo de los datos de la Web Social Semántica como una "cadena alimenticia" (ver Figura 2.5). Una cadena alimenticia en su contexto original "es el proceso de transferencia de energía alimenticia a través de una serie de organismos, en el que cada

uno se alimenta del precedente y es alimento del siguiente"[7]. Luego, en el contexto de la Web Social Semántica, se refiere al consumo de los datos desde la producción y recolección.

Se conoce como *productores* a los complementos o aplicaciones que permiten representar la información no semántica en un lenguaje semántico. Actualmente se han desarrollado complementos para diversos motores de foros, blogs y CMS que permiten la publicación de sus datos semánticamente. Por ejemplo el proyecto SIOC (Semantically-Interlinked Online Communities) cuenta con complementos para PHPBB2, Drupal, Wordpress, entre otros.

Se conoce como *recolectores* a aquellas aplicaciones que permiten integrar datos semánticos de diferentes fuentes, como por ejemplo los motores de búsqueda semánticos. Finalmente, se llama consumidores a aquellas aplicaciones que utilizan los datos generados por los productores y/o los datos agregados por recolectores. Dentro de los consumidores se encuentran los exploradores de sitios web sociales semánticos y las aplicaciones de análisis, por ejemplo: una aplicación que permita realizar análisis de redes sociales, una aplicación que permita visualizar la información agregada de dos foros, entre otros.

La Web Social Semántica podría enriquecer el estudio de sitios web sociales, ya que sería posible aplicar análisis de redes sociales sobre la información agregada de diferentes sitios web. Por ejemplo, se podrían buscar los expertos en un tema específico, integrando diferentes sitios relacionados entre sí [19]. Por otro lado, dado que la Web Social propone un estándar para la representación de los datos, sería posible generar algoritmos estándares y no dependiente del un modelo de datos específico, logrando un mayor de nivel de automatización y reutilización.

2.1.3.3. Friend-of-a-friend o FOAF

El proyecto FOAF comenzó el año 2000[8] y su objetivo es proveer un vocabulario interpretable para las máquinas que permita describir a la personas, las relaciones entre éstas y las cosas que ellas hacen o crean. FOAF provee una manera para interconectar los sitios web sociales utilizando tecnologías como RDF. FOAF también puede combinarse con otros vocabularios semánticos como SIOC, SKOS, etc.

Mediante FOAF las personas podrían portar y reutilizar sus datos en diferentes sitios web sociales. Actualmente existen algunos servicios de redes sociales que publican sus datos utilizando FOAF, como por ejemplo: hi5 o Vox.

[7] http://en.wikipedia.org/wiki/Food_chain [Fecha último acceso: marzo de 2011]
[8] http://www.foaf-project.org/about [Fecha último acceso: marzo de 2011]

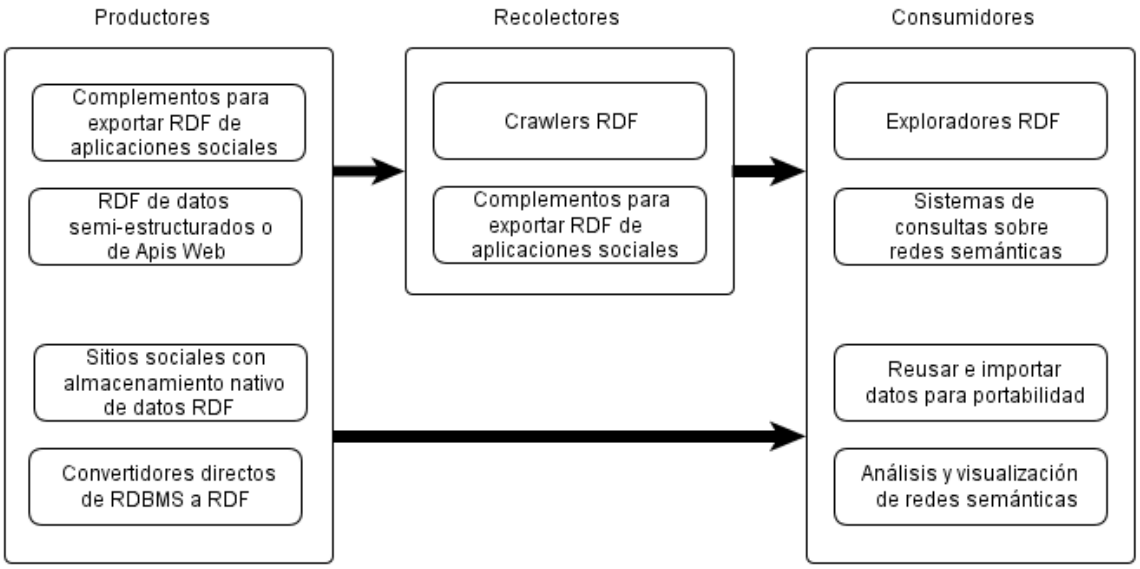


Figura 2.5: Cadena alimenticia Web Semántica

A continuación se muestra un ejemplo en el cual se utiliza FOAF para describir los datos de una persona llamada "Francisco Bustos", sus intereses y sus conocidos:

```
<rdf:RDF
xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:foaf="http://xmlns.com/foaf/0.1/"
xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#">
<foaf:Person>
<foaf:name>Francisco Bustos</foaf:name>
<foaf:mbox rdf:resource="mailto:fbustos@ing.uchile.cl"/>
<foaf:homepage rdf:resource="http://www.fbustos.cl/"/>
<foaf:nick>fbustos</foaf:nick>
<foaf:depiction rdf:resource="http://www.fbustos.cl/mi_foto.jpg"/
   >
<foaf:interest>
    <rdf:Description rdf:about="http://www.musica.com/"
       rdfs:label="Música" />
</foaf:interest>
<foaf:knows>
    <foaf:Person>
        <foaf:name>Angela Bustos</foaf:name>
        <foaf:mbox rdf:resource="mailto:angela@gmail.com" />
    </foaf:Person>
    <foaf:Person>
        <foaf:name>Sebastián Ríos</foaf:name>
        <foaf:mbox rdf:resource="mailto:sebastian@hotmail.com" />
```

```
    </foaf:Person>
</foaf:knows>
</foaf:Person>
</rdf:RDF>
```

2.1.3.4. Semantically Interlinked Online Communities o SIOC

El proyecto SIOC fue creado el 2004[9] y el 2007[10] fue reconocido por la W3C. Como su nombre lo dice, SIOC tiene por objetivo interconectar comunidades virtuales mediante de una representación semántica, como por ejemplo: blogs, foros y listas de correo. Al igual que FOAF; SIOC provee de un vocabulario semántico que puede ser procesable por máquinas.

La Figura 2.6 muestra la ontología o el modelo propuesto por SIOC, donde los rectángulos representan a recursos y las flechas representan propiedades de los recursos. Los objetos Role, UserAccount y Usergoup son los que permiten representar semánticamente la información de los miembros de un sitio y los permisos que éstos tienen. Forum y Post permiten representar semánticamente los sitios de discusión como foros, blog, listas de correos, etc. Forum tiene una referencia circular a sí mismo para representar semánticamente los foros agrupados en categorías. Post también tiene una referencia circular, lo que permite representar las respuestas de los tópicos. Tag/Category permite representar las etiquetas que los miembros de un sitio le asignan a los posts.

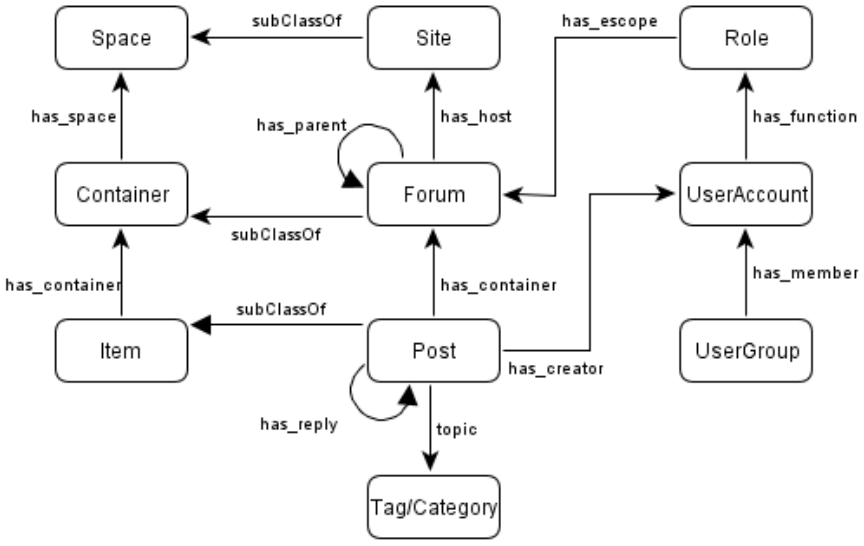


Figura 2.6: Ontología SIOC

[9]http://sioc-project.org/ [Fecha último acceso: marzo de 2011]
[10]http://www.w3.org/Submission/2007/02/ [Fecha último acceso: marzo de 2011]

A continuación se muestra un ejemplo de la representación semántica de un tópico utilizando SIOC:

```
<sioc:Post rdf:about="http://www.ejemplo.cl/post1">
<dcterms:title>"Hola, Soy nuevo en el foro"</dcterms:title>
<dcterms:created>2011-09-07</dcterms:created>
<sioc:has_container rdf:resource="http://www.ejemplo.cl/"/>
<sioc:has_creator>
    <sioc:UserAccount
      rdf:about="http://www.fbustos.cl" rdfs:label="fbustos">
    <rdfs:seeAlso rdf:resource="http://www.fbustos.cl"/>
</sioc:UserAccount>
</sioc:has_creator>
<sioc:content>"Hola, me gustaría saber acceder
al centro de descargas. gracias"</sioc:content>
<sioc:has_reply>
    <sioc:Post rdf:about="http://www.ejemplo.cl/post2">
       <rdfs:seeAlso rdf:resource="http://www.ejemplo.cl/post2"/>
    </sioc:Post>
</sioc:has_reply>
</sioc:Post>
```

Se puede desprender del texto que hay un sioc:Post, que fue creado por el usuario fbustos el 2011-09-07, que tiene por título "Hola, soy nuevo en el foro" y que tiene una respuesta identificada como http://www.ejemplo.cl/post2.

2.2. Análisis de redes sociales

Una red social es un conjunto de actores sociales (nodos o miembros) que están conectados entre sí por una o más tipos de relaciones. Los actores pueden ser individuos, grupos, organizaciones, páginas web, naciones, empresas, comunidades, entre otros. Las relaciones pueden ser cualquier cosa que relaciones a los actores, como por ejemplo: amistad, parentesco, amor, contacto por mensajes de un foro, entre otros.

El análisis de redes sociales (SNA) es una técnica basada en la teoría de redes que permite obtener información de la estructura de una red social a través de las relaciones existentes entre sus actores. En SNA una red social es modelada como un grafo, en donde sus actores son los nodos y las relaciones entre los actores son las aristas.

Los métodos tradicionales se basan en el estudio de los atributos de los actores, sin embargo, SNA permite estudiar las interacciones entre los actores y como éstos influyen en otros Scott [25].

Un sitio web social puede ser modelado como una red social, en donde sus miembros pueden ser vistos como actores. Las relaciones entre los miembros de un sitio web social surgen de la interacción de éstos a través de los medios de comunicación sociales. Por ejemplo: si A crea un post y B le responde, entonces A y B están relacionados socialmente. En ese sentido, el análisis de redes sociales (SNA) surge casi de manera natural para el estudio de los sitios web sociales.

2.2.1. Indicadores Básicos SNA

Algunos indicadores de SNA que sirven para describir y extraer información, a partir de la estructura de una red, son [25, 29]:

1. *Densidad*: La densidad describe el nivel general de acoplamiento o vinculación entre los nodos de un grafo. Un grafo "completo" es aquel en que cada nodo está conectado directamente con los demás. La densidad pretende medir cuan lejos se encuentra un grafo de su equivalente completo. Matemáticamente, la densidad se cálcula como el número total de aristas dividido por el número total de aristas posibles (grafo completo).

2. *Caminos y Distancias*: Dos nodos pueden estar conectados entre sí directamente o a través de otros nodos. Un camino es el conjunto de nodos y aristas que unen dos nodos cualquiera. En ese sentido, el largo de un camino equivale a la cantidad de aristas que forman dicho camino. La distancia entre dos nodos corresponde al camino más corto que los une.

3. *Centralidad*: La medida de centralidad permite identificar a los actores más importantes de un grafo. Los indicadores más usados son:

    a) *Centralidad de grado (degree centrality)*: La centralidad de grado corresponde al número de vecinos de un nodo cualquiera. Luego, un actor es importante si tiene un gran número de vecinos.

    b) *Centralidad de intermediación (betweenness centrality)*: La centralidad de intermediación es el número de veces que un nodo conecta a otro par de nodos. Este

indicador permite encontrar nodos intermediarios, los cuales son claves para la conectividad de la red.

c) *Centralidad de cercanía (closeness centrality)*:La centralidad de cercanía corresponde a la distancia promedio de un nodo al resto de los nodos. El actor con menor centralidad de cercanía es el más central o indepediente para alcanzar a otros nodos de la red.

4. *Clique*: Se llama clique a un sub-grafo, en el cual cada nodo está directamente conectado con los demás. Los clique sirven para detectar grupos o comunidades dentro de un grafo. Para la detección de comunidades también suele utilizarse algoritmos de clustering, como por ejemplo: k-means, algoritmos genéticos, entre otros.

La Figura 2.7 muestra un ejemplo de red social, cuyos actores son: A, B, C, D, E y F. Se puede desprender de la figura que existen dos cliques, los cuales son: {A, B, C} y {D, E, F}. La densidad del grafo es de 0.367, es decir, tiene un 37 % de las aritas del grafo completo. La distancia promedio entre dos nodos que se pueden alcanzar es de 1.667.

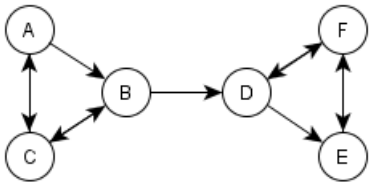


Figura 2.7: Ejemplo de red social

La Tabla 2.1 muestra las medidas de centralidad para el grafo de la Figura 2.7. Se puede despreder de la tabla que centralidad de grado (in-degree y out-degree) es casi homogénea, es decir, cada nodo tiene casi la misma cantidad de vecinos. A partir de las medidas de la centralidad de intermediación, se perciben 2 grandes intermediarios: B y D. En el caso de la centralidad de cercanía ocurre algo particular, D y F tienen centralidad igual a 1.0 debido a que no existe un camino desde éstos hacia el subgrafo {A, B, C}. El nodo con menor centralidad de cercanía promedio y que puede alcancar todos los nodos del grafo es B.

2.2.2. Hits

Hits es un algoritmo que fue diseñado por Kleinberg [17] para identificar los documentos (por ejemplo: páginas web) más importantes. Hits se basa en dos prinpicios básicos: Si un documento d es enlazado por muchos, entoces éste es importante; Y si los documentos

	in-degree	out-degree	centralidad de intermediación	centralidad de cercanía
A	1	2	0.0	2.0
B	2	2	6.0	1.6
C	2	2	1.0	2.0
D	2	2	6.0	1.0
E	2	1	0.0	1.5
F	2	2	1.0	1.0

Cuadro 2.1: Medidas de centralidad

que refieren a d ya son importantes, entonces d es aún más importante. A pesar que Hits fue diseñado para la clasificación de páginas web, se puede extender su uso para la indentificación de los actores más importante de una red social.

Hits se basa en el cálculo de dos indicadores: hub y authority (ver Figura 2.8). Sea un nodo p cualquiera, hub mide el grado de recomendador de p, y authority el grado de cuán buen recurso es p.

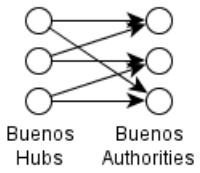


Figura 2.8: Authorities y hubs

Las ecuaciones 2.1 y 2.2 muestran como se calcula el grado de authority y hub de un nodo p, respectivamente. La notación $q \to p$ indica que q referencia a p (existe una arista). Luego, se puede desprender de ambas ecuaciones que el cálculo de un indicador depende de la sumatoria del otro, esto revela los principios fundamentales de Hits: un buen hub direcciona a muchos buenos authorities y un buen authority es apuntado por muchos buenos hubs.

$$x_p = \sum_{q, q \to p} y_q \tag{2.1}$$

$$y_p = \sum_{q, p \to q} y_p \tag{2.2}$$

2.2.3. PageRank

PageRank es una algoritmo para medir la importancia de una página web en Internet. Éste algoritmo fue diseñado y patentado por los fundadores de Google[11]. Si bien, el algoritmo actual de PageRank que utiliza Google se desconoce, en [7] se puede encontrar la definición inicial de PageRank.

La Ecuación (2.3) muestra el algoritmo inicial de PageRank. $PR(A)$ es el PageRank de la página A, el parámetro d es un factor de amortiguación que está entre 0 y 1 (usualmente se utiliza $d = 0,85$), $PR(i)$ es el PageRank de cada una de las páginas que apuntan a A y $C(i)$ es el número total de enlaces salientes de la página i.

$$PR(A) = (1 - d) + d * \sum_{i=1}^{n} \frac{PR(i)}{C(i)} \tag{2.3}$$

En Brin and Page [7] se da una justificacion intuitiva al algoritmo de PageRank, la que se basa en que los usuarios visitan los sitios web aleatoriamente. La probabilidad de que un usuario (random surfer) visite una página cualquiera corresponde a su PageRank. Desde este punto de vista, la Ecuación (2.3) se puede interpretar como: La probabilidad de que un usuario alcance una página (o PageRank) corresponde a la suma de las probabilidades de hacer click en los vínculos que llevan a dicha página. La probabilidad de que un usuario haga click en una página está dada por el número de vínculos a dicha página.

Sin embargo, los usuarios no hacen click indefinidamente. Llega un momento en que éstos pierden el interés por seguir navegando (haciendo clicks) y pueden saltarse a otra página aleatoriamente. El factor d pretende medir el interés del usuario por seguir haciando clicks en las páginas. Y por el contrario el factor $1 - d$ corresponde a la probabilidad de que el usuario deje de hacer clicks y se salté a otra página cualquiera.

A pesar que PageRank fue diseñado para medir la importancia de las páginas web, también se puede extender su uso para la clasificación de los actores de una red social.

[11] http://www.google.com/patents?vid=6285999 [Fecha último acceso: abril de 2011]

2.2.4. Configuración de la red

Construir el grafo, a partir de las interacciones entre los miembros de un sitio web social, no es una tarea sencilla. Para ilustrar mejor el problema se usará el ejemplo de la Figura 2.9, que representa un hilo de discusión del que participan 4 miembros (U1, U2, U3 y U4) y el cual es iniciado por U1 con el post 1. Es importante notar que el post 2 y el post 4 generan sub-hilos de discusión.

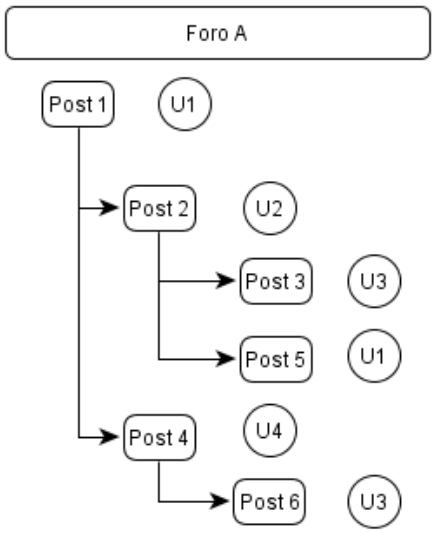


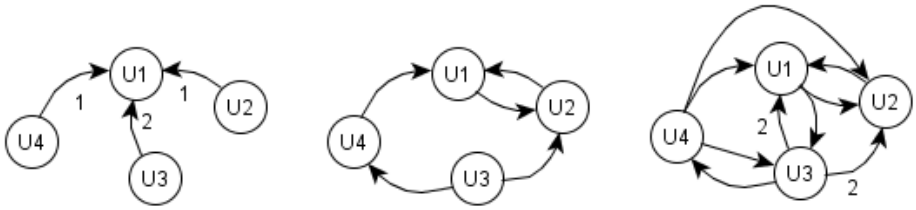
Figura 2.9: Ejemplo de un hilo de discusión

Al momento de construir la red, surge la siguiente interrogante: ¿El post 3 está dirigido al post 2 o al post 1? ¿O a ambos?. En términos generales no es trivial determinar si una respuesta está dirigida sólo al mensaje anterior, al autor del tópico, o a todos los mensajes anteriores. En la literatura ([19, 23, 2]) se pueden encontrar tres topologías para el estudio de sitios web sociales. Las topologías propuestas son: creador, última respuesta, y respuesta a todos. En la Figura 2.10 se muestra un ejemplo de estas tres topologías.

En la topología del creador todos los mensajes van dirigidos al creador del tema. En la topología de última respuesta cada mensaje va dirigido al mensaje anterior. Y en la topología de respuesta a todos, un mensaje va dirigido a todos los mensajes anteriores.

2.2.5. Mejorando SNA

Una limitante de SNA es que no considera el contenido de la interacción entre los usuarios. SNA no es capaz de filtrar los mensajes sin sentido u otra cosa que no corresponda



(a) Creador (b) Respuesta al anterior (c) Respuesta a todos

Figura 2.10: Ejemplo topologías creador, última respuesta y respuesta a todos

a un mensaje apropiado. Cualquier análisis relacionado con el contenido mismo que generan los usuarios, escapa del alcance de SNA. Sin embargo existen trabajos en que se utiliza SNA en combinación con minería de datos para reducir este problema. Minería de datos permite considerar el "significado de las relaciones" lo que permite construir redes más limpias. En [19] se implementan dos soluciones para mejorar SNA basadas en minería de datos, usando LDA (Latent Dirichlet Allocation) y CB (Concept-Based).

LDA permite extraer los principales tópicos (T) implícitos en un listado de documentos (D), para luego encontrar una distribución de probabilidad para cada documento d sobre el conjunto de tópicos T. LDA permite representar cada documento como un vector $d_i = (v_1, v_2, v_3, \ldots, v_n)$, donde cada v_i corresponde a la probabilidad de que el documento d se relacione con el tema t_i.

Por otro lado, mediante un número de conceptos generados por un experto del sistema de estudio, CB permite modelar cada documento como un vector $d_i = (c_1, c_2, c_3, \ldots, c_m)$, en que cada componente c_i equivale al peso que tiene d_i con respecto al concepto i.

En ambos casos, el contenido de la interacción entre los miembros de un sitio web social, puede ser modelado como un vector, cuya metodología de cálculo va depender si se utiliza CB o LDA. Luego, para considerar el contenido de las interacciones en el análisis de redes sociales se debe considerar la similitud entre vectores. Por ejemplo, para el caso de dos posts, se debe comparar el vector del posti con el vector del postj, si ambos vectores son parecidos, ambos posts también lo son. Y por el contrario, si los vectores son ortogonales, los posts también lo son.

2.3. ¿Qué es un framework?

En el contexto en que los recursos de dinero y tiempo son escasos, y en que las aplicaciones que se requieren en todo tipo de negocio son cada día más complejas, la reutilización de software surge casi de manera natural. La reutilización de software se conoce como el proceso de implementar aplicaciones reutilizando componentes de software ya existentes [28].

Hoy en día la reutilización ya no es una forma distinta de desarrollar, sino que es la única aproximación realista para llegar a los índices de productividad y calidad que la industria del software necesita [20]. En la literatura es común encontrar que las ventajas de la reutilización son mejor productividad (menor tiempo de desarrollo) y mayor calidad del software (mayor confiabilidad), sin embargo la reutilización puede hacerse en todos los niveles de desarrollo de software [25] (especificaciones, diseños, código, pruebas, documentación, etc.). A cada elemento de software reutilizable se le conoce como un activo de software.

Los frameworks han tenido bastante éxito como una forma de reutilización de software, debido a la diversidad de activos de software que permiten incorporar (arquitectura, diseño, know how de dominios específicos, código, etc.).

Se define un framework como un conjunto de patrones de diseño y componentes de software que permiten resolver problemas aplicados a un dominio específico de manera más rápida, y que justifica su desarrollo cuando se utiliza para la implementación de muchas aplicaciones [24]. También se les define como un conjunto de clases que envuelven un diseño abstracto de soluciones para una familia de problemas [14]. En ese sentido, se puede decir que un framework es un sistema incompleto, el cual se completa cuando es utilizado para el desarrollo de alguna aplicación específica.

2.3.1. Tipos de frameworks

Los framework se pueden clasificar de diversas maneras, sin embargo, las más conocidas son: la clasificación de acuerdo a su funcionalidad, y por la técnica en que son extendidos. A continuación se expondrán ambas clasificaciones.

Los frameworks de acuerdo a su funcionalidad se clasifican como frameworks de aplicación, de dominio y de soporte [26].

Frameworks de aplicación: Los frameworks de aplicación son aquellos que ofrecen soporte para el desarrollo de agún tipo de software o aplicacion, por ejemplo: un framework para el desarrollo de intercaces gráficas. Dichas funcionalidades usualmente involucran interfaces, documentos, base de datos, etc. Un ejemplo de este tipo de framework es JFC (Java Foundation Classes). JFC es un framework que permite construir interfaces gráficas de usuarios (GUIs) portables basadas en java.

Frameworks de Dominio: Estos frameworks ayudan a implementar programas en un dominio específico. El término framework de dominio es usado para denotar frameworks de un dominio específico. Un ejemplo de aplicación de dominio es un sistema de alarma de stock o sistema bancario. Los aplicaciones de dominio por lo general tienen que ser construídas a medida para una empresa y desarrolladas a partir de cero. Los frameworks ayudan a reducir el trabajo necesario para implementar aplicaciones en un dominio dado. Además permiten aumentar la calidad del software.

Frameworks de soporte: Los frameworks de soporte típicamente se desarrollan en dominios específicos relacionados con la computación, como el sistema de archivos o el administrador de memoria. Dar soporte para este tipo de dominios es necesario para simplificar el desarrollo de programas. Este tipo de frameworks es típicamente utilizado en conjunto con los frameworks de dominio y/o aplicación.

Los frameworks, de acuerdo a la técnica en que son extendidos, se clasifican como caja blanca, caja negra o una combinación de ambos [9].

Frameworks de Caja Blanca: Este tipo de frameworks se basa principalmente en la extensión por herencia y composición, es decir, la extensión de funcionalidades existentes se realiza creando subclases de una clase "mayor" existente en el framework (herencia) y sobrescribiendo métodos predefinidos. Este tipo de frameworks requiere que los desarrolladores de aplicaciones deban conocer cómo funciona y se compone el framework.

Frameworks de Caja Negra: Este tipo de frameworks es extendido a través de la composición y reodernamiento de componentes. Estos frameworks tienen la ventaja de que no se requiere conocer los detalles internos para su extensión. En general éste tipo de framework es más sencillo de usar y extender que los frameworks de caja blanca. Sin embargo son más difíciles de desarrollar ya que requiere que los diseñadores del framework se anticipen al amplio rango de uso del framework.

En general los frameworks presentan características de caja negra y blanca, y es muy díficil encontrar un framework que sea sólo de caja negra o caja blanca.

2.3.2. Arquitectura: hotspots y frozenpots

Un requisito principal en el diseño de frameworks es que sean flexibles para ser adaptados a cada aplicación en específico, a esta propiedad se le conoce como *hotspots*. Hay dos tipos de hotspots dependiendo del tipo de framework (caja blanca o caja negra). Los hotspots en el caso de frameworks de caja blanca son las clases o métodos abstractos que deben ser implementados, mientras los hotspots en los frameworks de caja negra consisten en el ordenamiento de los componentes ya existentes mediante archivos de configuración[26].

Se conooce como *frozenspots* a aquellos elementos que definen la arquitectura de un sistema, es decir, son los componentes básicos de éste (núcleo del framework). A diferencia de los hotspots, los frozenspots no cambian en la instanciación del framework.

Los frozenpots y hotsposts corresponden a conceptos referentes a la arquitectura de un framework.

2.3.3. El proceso de desarrollo de frameworks

El proceso de desarrollo de frameworks es más complejo que el desarrollo de una aplicación, ya que los diseñadores deben tomar decisiones de diseño para resolver una familia de problemas y los usuarios del framework deben comprender dichas decisiones. Por eso es importante tener en cuenta las siguientes buenas práticas.

En [13] se explica la forma "ideal" y "buena" de desarrollar un framework. La forma ideal se basa en las tres etapas. Primero, analizar el dominio del problema y recolección de ejemplos a construir. Segundo, diseñar abstracciones que puedan cubrir todos los ejemplos. Y último, probar el framework resolviendo los ejemplos. La forma ideal es imposible de seguir porque es muy difícil y costoso hacer un análisis profundo de un dominio específico.

La forma buena para desarrollar frameworks es un poco más flexible y menos costosa, la que propone construir un framework en base a dos aplicaciones tipo.

Otra manera de desarrollar framework y complementaria a la vez, es la que se propone en [5], donde se distiguen dos actividades en el proceso de desarrollo de frameworks: diseño del centro del framework y desarrollo de incrementos internos. El diseño del centro del framework involucra las clases concretas y abstractas. Las clases concretas son aquellas que son propias del framework y que el usuario nos las puede manipular, y las clases abstractas son aquellas que pueden usarse mediante la creación de subclases.

Los incrementos internos del framework correspoden a un conjunto de clases que forman librerías. Dichas clases se encargan de implementar comportamientos típicos del framework, es decir, corresponden a una instaciación específica del framework.

A continuación se mencionan algunos requerimientos deseables de un framework:

Completo: Un framework debe proveer la mayor cantidad de funcionalidades que se requieran para el dominio en que fueron creados. Además debe proveer de funcionalidades por defecto dentro de lo posible.

Flexible: Un framework debe proveer de abstracciones que puedan utilizarse en diferentes contextos y problemas.

Extensible: Los usuarios del framework deben poder agregar y modificar funcionalidades fácilmente.

Comprensible: Este es unos de los requerimientos fundamentales de un framework, ya que si éste no es comprensible, entoces no se usará. Para que un framework sea compresible es necesario que existan ejemplos y una buena documentación, además de seguir estándares de desarrollo.

Capítulo 3

Especificación del Problema y Descripción de la Solución

3.1. Especificación del Problema

Actualmente existen diversos incentivos para aplicar la Web Semántica a la Web Social. Entre estos se pueden mencionar:(1) interconectar los sitios web sociales, que hoy conviven aislados, (2) proveer las condiciones necesarias para la portabilidad de los datos de los usuarios, (3) generar las condiciones necesarias para crear nuevos servicios en base a un lenguaje interpretable por máquinas, y por último (4) simplificar y mejorar el estudio de sitios web sociales basado en análisis de redes.

No obstante, desarrollar herramientas que permitan llevar sitios web sociales a la Web Semántica no es una tarea sencilla. El proyecto SIOC (descrito en la sección 2.1.3.4) ha intentando cubrir todas las etapas de la cadena alimenticia de la Web Social Semántica (producción, recolección y consumo de datos semánticos) mediante el desarrollo de aplicaciones como exportadores RDF, exploradores RDF y recolectores RDF. Por ejemplo, dentro de los productores están los exportadores para PHPBB2, Drupal, WordPress, entre otros. Sin embargo, se pueden mencionar los siguientes problemas de los exportadores de SIOC:

- Para crear un exportador se requiere conocer en profundidad la arquitectura del sitio web a exportar.

- Si bien todos los exportadores se basan en la ontología de SIOC, no siempre generan grafos RDFs (archivos RDF) equivalentes.

- En general los exportadores van embebidos como un complemento en los sitios web, por lo tanto se necesita generar un API para cada lenguaje de programación.

- Los exportadores en general se limitan a exportar un sólo formato de archivo, sin dar la posibilidad de utilizar otros formatos, como por ejemplo N3 o N-Triples.

- Los exportadores generan los datos a medida que son requeridos por un recolector (régimen perezoso), lo que garantiza un mejor desempeño para la navegación de los datos, pero dificulta la recolección éstos para su análisis.

Considerando los problemas mencionados previamente, el framework desarrollado debiese cumplir los siguientes requerimientos de software:

Una API para generar SIOC: El framework debe ser capaz de proveer una API que permita encapsular la generación y exportación de datos RDFs. Como entrada debe recibir un conjunto de objetos y como salida debe retornar archivos RDF basados en la ontología SIOC. La API tiene como propósito facilitar el desarrollo de exportadores RDF.

Además, la API debe proveer la generación de datos en régimen perezoso y en régimen ansioso. En régimen perezoso, los datos semánticos se generan a medida que son requeridos. En régimen ansioso, todos los datos son generados de una sola vez. Este último es necesario para poder realizar análisis de los foros web.

Un Complemento para persistir los datos semánticos: Si bien la API debe ser capaz de exportar los datos RDFs a archivos, el framework debe proveer de un módulo que permita persistir los datos semánticos en una base de datos convencional y en un sistema de archivos. Este requerimiento está pensado para facilitar el análisis de los foros web.

Un exportador de PHPBB2 a SIOC: Para completar la etapa de la producción de datos, el framework debe proveer de un exportador para el motor de foros PHPBB2. Este exportador será clave para el futuro desarrollo de exportadores.

Un Exportador de grafos para SNA: El framework debe ser capaz, a partir de los datos semánticos, generar las tres topologías de red descritas en la sección 2.2 (creador, respuesta al anterior, respuesta a todos los anteriores). Además, el exportador SNA debe ser capaz de generar grafos a partir de la información agregada de varios sitios web (interoperabilidad).

Los grafos generados por el exportador SNA, deben ser exportados como archivos

pajek, el cual es el formato más usado por los programas de análisis de redes sociales.

3.2. Descripción de la Solución

La Figura 3.1 muestra un diseño global del framework, en base a la cadena alimenticia de la Web Social Semántica. El framework se centró en la etapa de la producción de datos y el consumo directo de éstos. Dentro de la producción se desarrolló una API y el exportador para PHPBB2. Y dentro de los consumidores se desarrolló el exportador de grafos SNA.

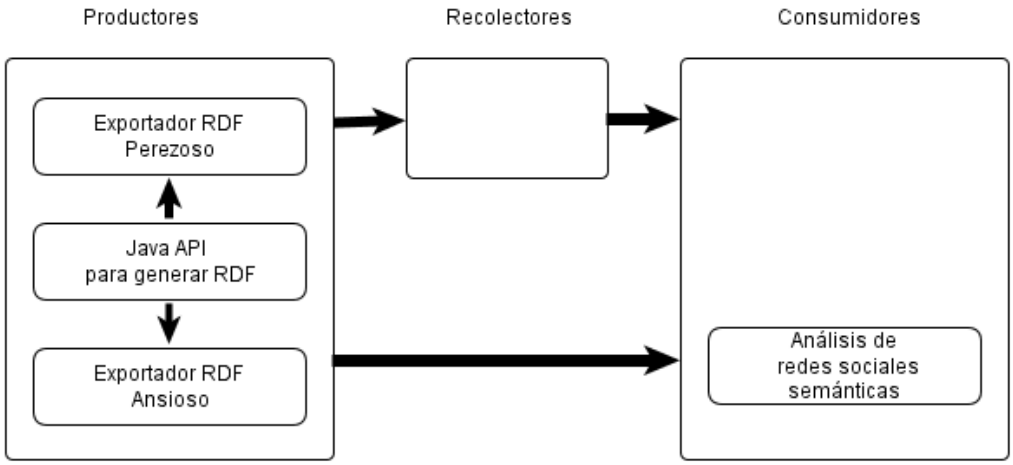


Figura 3.1: Cadena alimenticia framework

En la sección 2.1.3.4 se expusieron los fundamentos de la ontología de SIOC. Además, en la sección 2.2 se explicaron los beneficios de realizar análisis de redes sociales considerando el significado de las interacciones entre los miembros de un sitio, mediante técnicas de minería de datos como LDA.

Luego, como resultado de aplicar LDA a un foro web, se obtiene un vector de probabilidad para cada post, en que la componente i equivale a la probabilidad de que un post se relacione con el término i. Dichos vectores sirven para comparar la semejanza entre los mensajes de una red social.

En ese sentido se generó una extensión de la ontología SIOC[1], para poder almacenar los vectores de probabilidad (LDA) o de pesos (CB). La extensión básicamente se basa en

[1]Trabajo desarrollado en conjunto con el equipo de inteligencia Web de la Universidad de Chile, http://wi.dii.uchile.cl/ [Fecha último acceso: marzo de 2011]

agregar un nuevo objeto a SIOC, llamado Concept, como lo muestra la Figura 3.2. El predicado o propiedad que permite vincular a Post con Concept es hasConcept. Además se crearon las siguientes propiedades para Concept: sioc:conceptId (identificador del concepto o término), rdfs:label (nombre del concepto o término) y sioc:weight_value (peso o probabilidad).

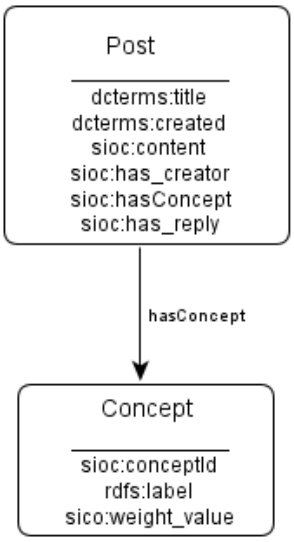


Figura 3.2: Extensión SIOC

A continuación se muestra un ejemplo de un post representado semánticamente, utilizando SIOC Extendido (la letra “a” equivale a “rdf:type”).

```
<http://www.ejemplo.cl/post1> a sioc:post ;
dcterms:created 1259943235000 ;
dcterms:title "Ayuda"@es ;
sioc:content "Hola, soy nuevo en foro... necesito ayuda"@es ;
sioc:hasConcept
    [ a sioc:Concept ;
    rdfs:label "término 1" ;
    sioc:conceptId 1 ;
    sioc:weight_value "0,314" ] ;
sioc:hasConcept
    [ a sioc:Concept ;
    rdfs:label "término 2" ;
    sioc:conceptId 2 ;
    sioc:weight_value "0" ] ;
sioc:hasConcept
    [ a sioc:Concept ;
    rdfs:label "término 3" ;
    sioc:conceptId 3 ; sioc:weight_value "0,8" ] ;
sioc:has_creator <http://www.fbustos.cim> ;
```

```
sioc:has_reply <http://www.ejemplo.cl/post3>.
```

A parte de los datos propios de un post (título, contenido y respuestas), se observa la presencia del vector de concepto, que se compone de tres términos. Se puede desprender que el post tiene poca relación con el término 2, debido a que tiene probabilidad 0. Sin embargo, el post tiene mucha relación con el término 3, ya que en ese caso tiene probabilidad 0,8.

El framework desarrollado utilizó como base el modelo SIOC extendido, sin embargo cabe señalar que la implementación de las técnicas LDA y Concept-Base no formaron parte de este trabajo. Para la validación del framework, se asumió que se contaba con los vectores (o matriz) de conceptos. No obstante, sería interesante abordar el desarrollo de técnicas como LDA y CB, en base a una representación semántica de los sitios web sociales.

3.2.1. Una API para generar SIOC

La Figura 3.3 muestra una estructura general de un foro web. Un foro web (Sitio) puede contener foros y usuarios. A su vez los foros pueden contener otros foros. Los posts pueden contener otros posts (respuestas). Los usuarios son los que crean los posts.

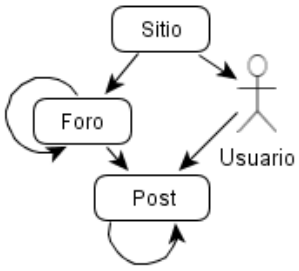


Figura 3.3: Esquema global foros web

Luego a partir de la Figura 3.3 se diseño la API. En primera instancia, se diseñó y desarrolló una API de régimen perezoso, la que genera datos semánticos a medida que son requeridos por una aplicación externa. Y luego se extendió la API perezosa para generar una API ansiosa, es decir, la que permite generar todos los datos de una sola vez. Cabe mencionar que para el desarrollo de la API también se consideró el vector de conceptos (vector generado por LDA o CB).

La Figura 3.4 muestra el diagrama de clase de la API con las principales clases, atributos y métodos. Las clases cuyo prefijo es “Lazy”corresponden a los objetos de la API perezosa y las clases cuyo prefijo es “Eager” corresponden a los objetos de la API ansiosa,

el prefijo hace referencia al régimen de la API. Se puede desprender de la figura que cada clase Eager extiende a su correspondiente Lazy, por ejemplo: la clase EagerPost extiende a LazyPost.

En los siguientes apartados, se realiza una breve descripción de cada API (perezosa o ansiosa) y sus principales clases.

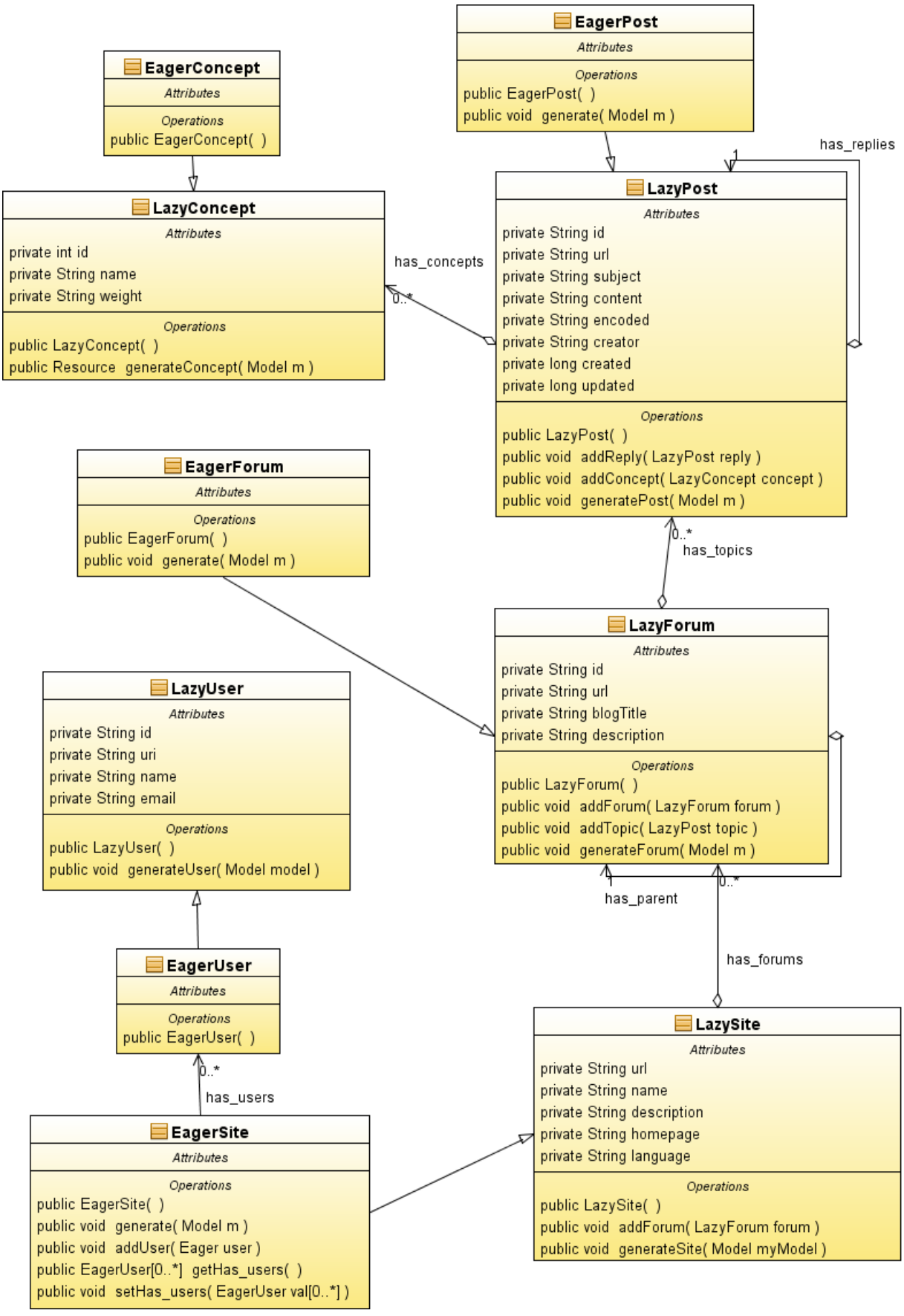


Figura 3.4: Diagrama de clases API

3.2.1.1. API perezosa

Las clases de la API perezosa son: LazyPost, LazyUser, LazyForum, LazySite y LazyConcept. Cada clase de esta API, mediante la instanciación, permite llevar un elemento (post, usuario, foro) de un foro web a RDF. Cabe mencionar que, la clase LazySite permite representar los metadatos del foro web a RDF. La clase LazyConcept permite representar los vectores de probabilidades (LDA) o de pesos (CB) en RDF en base a la extensión de SIOC propuesta en este trabajo.

Las relaciones entre las clases de la API perezosa se describen a continuación:

has_replies: Esta relación representa el vínculo entre un mensaje y sus repuestas. Los mensajes pueden tener muchas respuestas y a su vez, cada respuesta puede tener más respuestas. Para modelar este fenómeno, se agregó como atributo a la clase LazyPost una lista de <LazyPost>. El método addReply de la clase LazyPost permite asignar respuestas a un post, donde cada respuesta es también un post.

has_concepts: Esta relación permite asociar un post a un vector de conceptos, que puede ser generado por LDA o CB. El método addConcept de la clase LazyPost permite agregar un concepto a un post. Luego, el vector de conceptos se forma por todos los conceptos agregados al post.

has_topics: Un foro (o hilo de discusión) puede contener muchos mensajes. Luego, para modelar esta relación se agregó como atributo de LazyForum una lista de <LazyPost>. Mediante el método addTopic es posible agregar a un foro un post.

has_parent: Es común encontrar que los foros web están agrupados por categorías, las cuales también pueden ser modeladas por la clase LazyForum. Por lo tanto, un LazyForum puede contener otros LazyForum. Luego, para asignar un LazyForum a otro, se debe utilizar el método addForum.

has_forums: Se mencionó anteriormente que la clase LazySite permite representar los metadatos de un foro web. Un posible metadato es una categoría o foro padre, por lo tanto, un LazySite puede contener muchos LazyForum. Luego, mediante el método addForum de la clase LazySite es posible asignar un LazyForum a un LazySite.

Para la generación de los archivos RDF se utilizó la librería Jena[2] (versión 2.6.4), la cual fue creada para el desarrollo de aplicaciones semánticas. Jena provee de funcionalidades

[2]http://jena.sourceforge.net/index.html [Fecha último acceso: marzo de 2011]

como la generación de archivos RDF en RDF/XML, N3 y N-Triples, un motor de consultas SPARQL y otras más. Jena Provee de un objeto llamado Model (grafo RDF), sobre el cual se crean los recursos. La clase Model almacena las referencias a todos los recursos del grafo RDF, luego mediante un par de instrucciones es posible exportar el grafo RDF a un archivo RDF/XML, N3 o N-Triples.

Los métodos generatePost, generateUser, generateForum, generateSite y generateConcept permiten mapear los objetos de la API a un grafo RDF. Todos los métodos mencionados anteriormente, reciben como parámetro una instancia de la clase Model, sobre la cual se carga el grafo RDF.

Una de las ventaja de utilizar Jena, es que todo el RDF que genera es validado por un componente interno para que éste cumpla con la especificación de la Web Semántica.

3.2.1.2. API ansiosa

Las clases de la API ansiosa son: EagerPost, EagerUser, EagerForum y EagerConcept. Esta API corresponde a una extensión de la API perezosa y se caracteriza por generar el grafo RDF completo de una sola vez. Para generar el grafo RDF completo sólo se necesita llamar al método generate de la clase EagerSite, ya que éste invoca la generación de los foros, usuarios, posts y conceptos.

La relación has_users permite vincular todos los EagerUser de un foro al EagerSite. El método addUser sirve para agregar un EagerUser a un EagerSite. A continuación se muestra un ejemplo de uso de la API ansiosa:

```
//Se instancia EagerSite
EagerSite site = new EagerSite(http://plexilandia.cl/foro,
   Plexilandia,null,http://plexilandia);
// se crea un foro
EagerForum forum = new EagerForum(1, http://plexilandia.cl/foro1,
   primer foro, este es un foro para nuevos usuarios);
// se crea un tópico
EagerPost topic = new EagerPost(1, http://plexilandia.cl/post1,
   feliz navidad, saludos, "utf8", fbustos, 2009-12-25, null);
// se crea una respuesta
EagerPost reply = new EagerPost(2, http://plexilandia.cl/post2,
   null,gracias, utf8, sríos, 2009-12-26, null);
```

```
// se agrega la respuesta al tópico
topic.addReply(reply);
// se agrega el tópico al foro
forum.addTopic(topic);
// se agrega el foro al sitio
site.addForum(forum);
// se crea un usuario
EagerUser u = new EagerUser(1, http://www.fbustos.com,fbustos,
   fbustos@dii.cl);
// se agrega el usuario al sitio
site.addUser(u);

// Luego, para generar el RDF se ejecuta:
site.generate(model);
```

3.2.2. Un Complemento para persistir los datos semánticos

Jena provee dos formas distintas de persistir los datos semánticos; SDB[3] y TDB[4]. TDB permite persistir los datos en un sistema de archivos y SDB en una base de datos relacional. Sin embargo TDB es más eficiente y requiere de menos configuración.

Luego, para que el framework pudiese utilizar SDB o TDB indistintamente, se creó una clase que permitió encapsular (como una caja negra) las diferencias de utilizar entre dichos sistemas de persistencia . Dicha clase se llamó GenericModel. La Figura 3.5 muestra el esquema de la clase GenericModel, donde se puede apreciar los diferentes métodos para SDB y TDB. Cabe señalar que el uso de GenericModel es transversal al framework, o en otras palabras, GenericModel corresponde a un hotspot.

El impacto de la implementación de GenericModel sobre la API no fue tan grande. Básicamente, se tuvo que cambiar el enbezado de los métodos que generan RDF, donde decía "Model m" se tuvo que cambiar por "GenericModel m".

GenericModel ofrece la posibilidad de persistir los datos o simplemente trabajar con ellos en memoria. Si no se desea persistir los datos se debe setear la variable isInMemory a verdadero mediante el método setIsInMemory. Por el contrario, si de sea persistir los datos; se debe optar por TDB o SDB, por ejemplo: Para utilizar el sistema TDB (Sistema de archivos), se debe setear la variable isTDB a verdadero mediante el método setIsTDB. El método setTdbPath permite setear la ruta donde se persisten los datos.

[3]http://openjena.org/SDB/ [Fecha último acceso: marzo de 2011]
[4]http://openjena.org/TDB/ [Fecha último acceso: marzo de 2011]

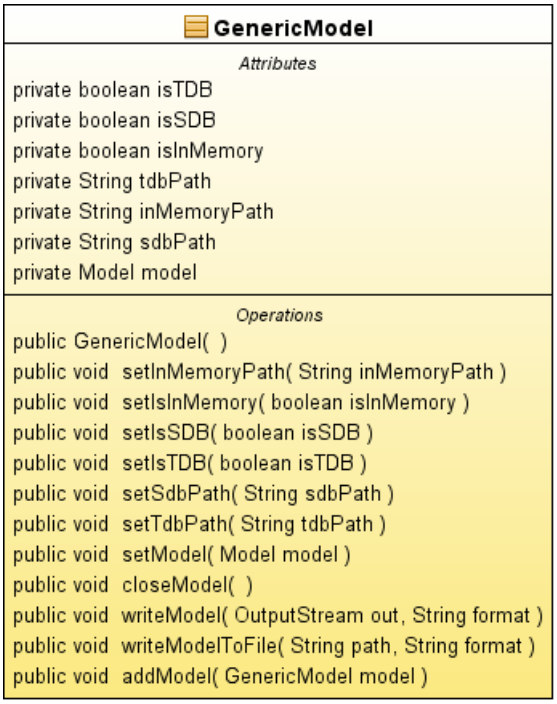


Figura 3.5: Esquema GenericModel

Utilizar SDB requiere un mayor nivel de configuración. Previamente a utilizar SDB, se debe configurar la base de datos a utilizar, en el anexo A se explican los pasos a seguir para configurar una base de datos para SDB en MySQL y Windows. Una vez configurada la base de datos, se debe setear la variable isSDB a verdadero. Mediante el método setSdbPath se debe setear la ruta del archivo que contiene los datos de conexión a la base de datos SDB. Cabe mencionar, que el archivo de configuración de los datos de conexión se debe construir en formato RDF, A continuación se muestra un ejemplo:

```
@prefix sdb: <http://jena.hpl.hp.com/2007/sdb#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix ja: <http://jena.hpl.hp.com/2005/11/Assembler#> .
# MySQL - InnoDB
<#store>
rdf:type sdb:Store ;
sdb:layout "layout2" ;
sdb:connection <#conn> ;
sdb:engine "InnoDB" ;
# MySQL.
<#conn>
rdf:type sdb:SDBConnection ;
sdb:sdbType "MySQL" ;
# Necesarios para la conexión JDBC
```

```
sdb:sdbHost "localhost" ;
sdb:sdbName "sdb2" ;
sdb:sdbUser "root" ;
sdb:sdbPassword "root" ;
sdb:driver "com.mysql.jdbc.Driver" ; .
```

Finalmente, los métodos writeModel y writeModelToFile permiten exportar el grafo RDF a un archivo N3, RDF/XML o N-Triples. A continuación se muestra un ejemplo de uso de GenericModel.

```
// se instancia GenericModel
GenericModel model = new GenericModel();
//se setea TDB
model.setIsTDB(true);
model.setTdbPath("Escritorio/Plexilandia/tdb");
/*** Aquí se debe poblar el grafo RDF ***/
/* Se cierra la instancia (esto permite cerrar conexiones y
   sincronizar los datos que están en memoria con algún sistema
   de persistencia) */
model.closeModel();
/*Se exporta el grafo RDF a un archivo*/
model.writeModelToFile("plexilandia.n3", "N3");
```

Jena provee de funcionalidades que permiten realizar operaciones sobre los grafos RDF, como por ejemplo: unión, diferencia e intersección de grafos. El método addModel de la clase GenericModel, en base a la unión de modelos, permite agregar datos de diferentes modelos en uno sólo. Esta funcionalidad es clave para la integración de sitios web sociales.

3.2.3. Un exportador de PHPBB2 a SIOC

Para el desarrollo de este requerimiento fue necesario comprender el modelo de datos utilizado por PHPBB2. La Figura 3.6 muestra un extracto del modelo de datos de PHPBB2 con las principales tablas y atributos. Cabe mencionar, que los tipos de datos de los atributos no representan necesariamente los reales y que las relaciones entre las tablas fueron insertadas a próposito.

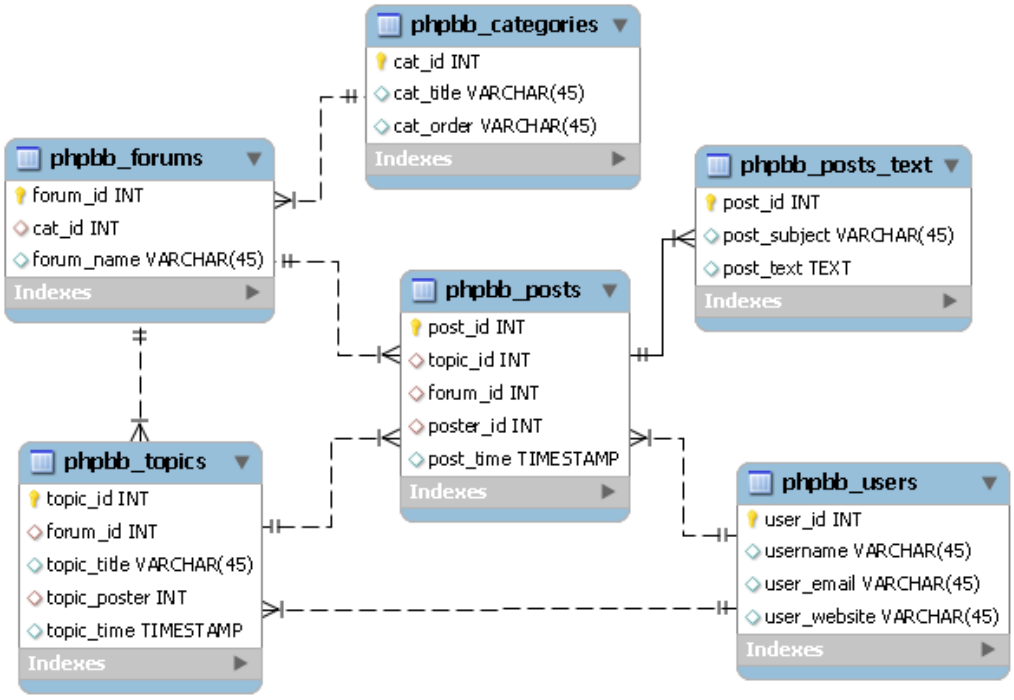


Figura 3.6: Modelo de datos de PHPBB2

La tabla phpbb_categories permite agrupar los foros. La información de los foros se almacena en phpbb_forums. Un foro puede tener muchos tópicos (primer mensaje de un hilo de discusión), lo cuales son almacenados en la tabla phpbb_topics. La tabla phpbb_posts permite almacenar el idenficador del autor de cada post (poster_id) y la fecha/hora de su creación. En la tabla phpbb_users se guarda toda la información referente a los miembros del foro y en la tabla phpbb_posts_text se almacena el título y el contenido de cada post.

Luego, los pasos necesarios para llevar un foro PHPBB2 a la Web Semántica, utilizando la API ansiosa, son:

1. Seleccionar todas las categorías.

2. Para cada categoría, obtener todos sus foros.

3. Para cada foro, obtener todos sus tópicos.

4. Para cada tópico, obtener todas sus repuestas.

5. Finalmente, obtener todos los miembros del foro.

En base a los pasos mencionados anteriormente, se contruyó el exportador de PHPBB2 a

SIOC. En el Anexo B se muestra un extracto de código del exportador, que utiliza los pasos mencionados anteriormente para cargar la API y generar el grafo RDF.

A continuación se muestra un ejemplo de uso del exportador:

```
/* se crea una instancia del exportador.El parámetro que recibe
    la clase Exporter corresponde a la ruta de un archivo de texto
     que contiene los datos de conexión a la base  de datos y el
    nombre de las tablas utilizadas en las consultas */
Exporter ex = new Exporter("src/plexilandia/config.properties");
// Método que carga los objetos de la API
ex.export();
/* Método que carga el grafo RDF en una instancia de GenericModel
     */
ex.generate(model);
```

En el Anexo C se muestra un ejemplo de archivo de configuración utilizado por el exporter.

La metodología utilizada para el desarrollo de este requerimiento se puede aplicar al desarrollo de otros exportadores, como por ejemplo: exportador para VBulletin. Inicialmente, el framework cuenta con un exportador para PHPBB2, pero se espera que en el futuro vaya creciendo con el aporte de la comunidad.

3.2.4. Un Exportador de grafos para SNA

Para el desarrollo de este requerimiento se utilizó la librería Jung[5] (Java Universal Network/Graph Framework), versión 2.0.1. Jung es una librería que provee de un lenguaje extensible para modelar, analizar y visualizar redes. Sin embargo en este requerimiento sólo se utilizó el modulo de construcción de redes para generar los archivos pajek (archivos utilizados para representar la información de una red) a partir de los datos semánticos de los foros.

El objetivo de este requerimiento no es la generación de indicadores de SNA, sino crear mecanismos para aplicar análisis de redes sociales sobre la estructura semántica generada por el framework. Luego, con la generación de los grafos y su exportación en archivos pajek se cumple con el objetivo de este requerimiento. Sin perjuicio de lo cual, este trabajo abre las posibilidades para automatizar la extracción de indicadores que permitan operar sobre sitios web sociales.

[5]http://jung.sourceforge.net/ [Fecha último acceso: marzo de 2011]

La Figura 3.7 muestra el diagrama de clases del exportador de grafos SNA. La clase principal de este requerimiento es NetExporter, la cual permite generar las tres topologías de redes estudiadas en la sección 2.2 (creador, repuesta al anterior, respuesta a todos los anteriores) mediante consultas SPARQL. La clase GenericVertex y GenericEdge permiten modelar los vértices y aristas de los grafos SNA, respectivamente.

Cabe Señalar, que la clase GenericEdge está compuesta por una lista de GenericArc, clase que sirve para almacenar las distancias entre los vectores y el módulo de cada vector. El tamaño de la lista de GenericArc de un arista cualquiera, es igual al número de relaciones que existen entre los nodos que forman la arista. La lista de GenericArc sirve para calcular el peso de las aristas de un grafo.

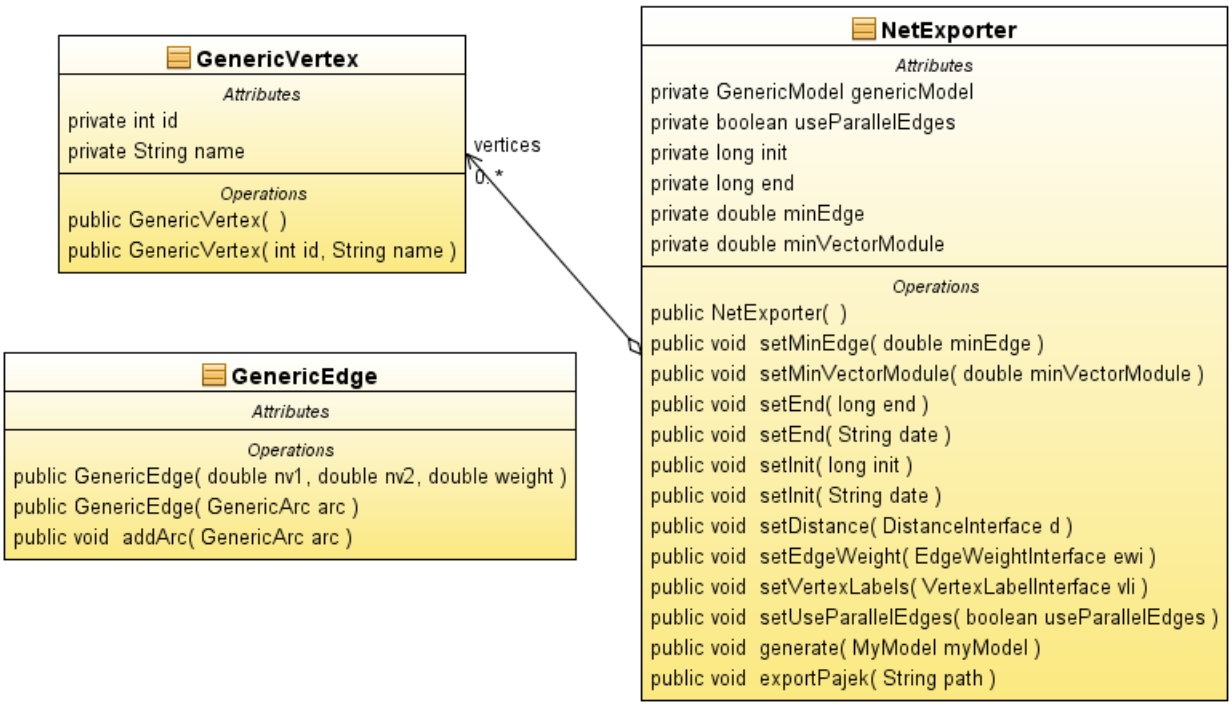


Figura 3.7: Diagrama de clases del exportador SNA

La clase NetExporter cuenta con métodos para: setear un periodo particular de estudio (setInit y setEnd), definir la función de distancia para la comparación de vectores (setDistance), definir el método de cálculo de los pesos de las aristas (setEdgeWeight), entre otros. A continuación se muestra un ejemplo de uso del exportador SNA:

```
//se incializa el exportador para SNA.
NetExporter ne = new NetExporter();
```

```
/* se generan los grafos SNA para el RDF embebido en model
   */
ne.generate(model);
//Se exportan los archivos PAJEK
ne.exportPajek("Escritorio/Plexilandia/graphs");
```

Sin embargo, la construcción de la red no es una tarea sencilla. No es posible encapsular todas las funcionalidades para crear un framework de caja negra. Existen 2 hotspots que son críticos para la configuración de la red. En las secciones 3.2.4.1 y 3.2.4.2 se explican en detalle.

3.2.4.1. Función de distancia entre vectores

Mediante las técnicas de CB y LDA, es posible representar los posts como vectores. Luego, para comparar el contenido entre dos posts es necesario comparar vectores. La función de distancia entre vectores varía entre aplicaciones, para ello se creó una interfaz llamada DistanceInterface (ver Figura 3.8).

La interfaz DistanceInterface contiene sólo un método, que recibe dos listas de números reales y entrega como resultado un GenericArc. Este objeto fue creado con la intención de almacenar el módulo de ambos vectores y la distancia o similitud obtenida entre ellos, para que en otros módulos del exportador se puedan aplicar filtros sobre estos valores.

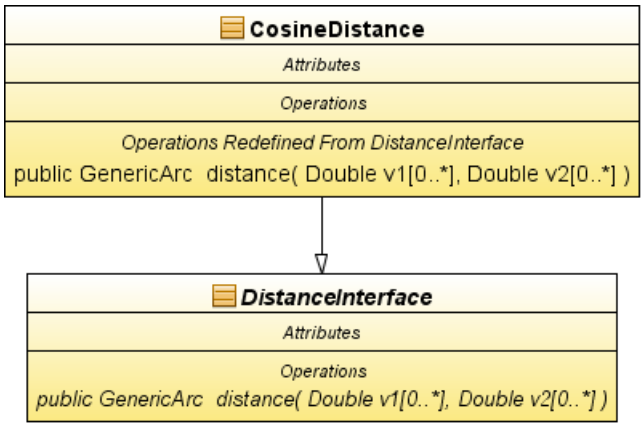


Figura 3.8: Diagrama de clases DistanceInterface.

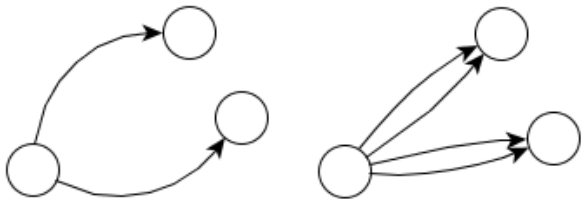
La clase CosineDistance utiliza el coseno (proyección de un vector en otro) como medida de similitud. Esta clase es utilizada por defecto por el framework. El método de cálculo utilizado por CosineDistance se muestra en la Ecuación (3.1). P_x corresponde al post del

usuario x, g_{xk} corresponde al componente k del vector P_x y cada vector es de K dimensiones (con $x = \{i, j\}$)

$$d(P_i, P_j) = \frac{\sum_k g_{ik} g_{jk}}{\sqrt{\sum_k g_{ik}^2 \sum_k g_{jk}^2}} \tag{3.1}$$

3.2.4.2. Cálculo del peso de las aristas

El framework debe permitir la generación de los grafos SNA con o sin aristas paralelas. Se conoce como aristas paralelas cuando existe más de una arista y con la misma dirección entre el mismo par de nodos de un grafo (ver Figura 3.9).



(a) Aristas normales (b) Aristas paralelas

Figura 3.9: Tipos de aristas

Algunas alternativas para reducir podría ser la suma de los pesos, el promedio, etc. Para generar este grado de libertad, se implementó la interfaz EdgeWeightInterface (ver Figura 3.10) que permite definir la operación a utilizar para reducir las aristas paralelas. El framework por defecto utiliza el promedio (clase EdgeWeight).

Otra funcionalidad de este hotspost, es la capacidad de filtrar las aristas de acuerdo a un criterio definido en base al peso de éstas y/o al módulo de los vectores de conceptos, como por ejemplo: no considerar las aristas cuyo peso sea inferior a 0,8.

La clase EdgeWeightFiltered permite reducir las aristas paralelas en base al promedio de los pesos y filtrar las aristas cuyo peso o cuyos módulos de vectores se encuentren bajo un umbral (minEdge y minVectorModule, respectivamente). Para activar el uso de esta clase se debe setear las variables minEdge y minVectorModule de la clase NetExporter, mediante los métodos setMinEdge y setMinVectorModule.

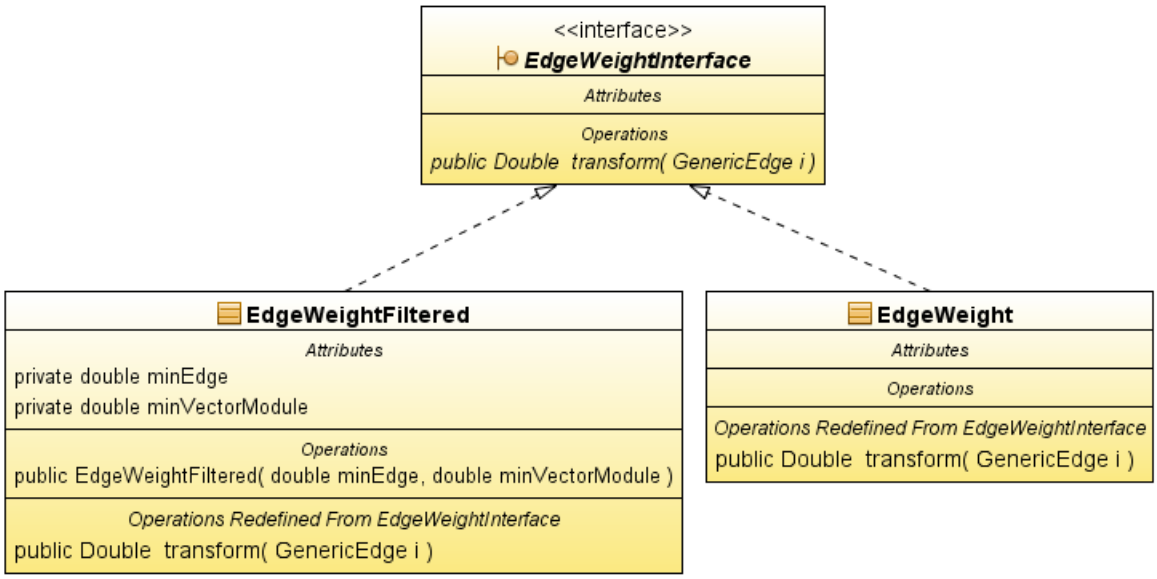


Figura 3.10: Diagrama de clases para el cálculo de pesos.

3.2.5. Ejemplo de uso del framework

A continuación se muestra un ejemplo sencillo de uso del framework.

```
public class Plexilandia {
public static void main(String[] args) throws SQLException,
IOException {
//Se incializa GenericModel
GenericModel model = new GenericModel();
model.setIsTDB(true);
model.setTdbPath("Escritorio/Plexilandia/tdb");
//Se incializa el exportador
Exporter ex = new Exporter("src/plexilandia/config.properties");
//Se carga la API
ex.export();
// se carga el grafo RDF
ex.generate(model, null);
// se genera el archivo RDF
myModel.writeModelToFile("Escritorio/Plexilandia/plexilandia.n3",
   "N3");
//se incializa el exportador para SNA.
NetExporter ne = new NetExporter();
// se define el periodo a estudiar
ne.setInit("2009-06-01 00:00:00");
```

```
ne.setEnd("2009-06-30 00:00:00");
// se generan los grafos SNA
ne.generate(model);
//Se exportan los archivos PAJEK
ne.exportPajek("Escritorio/Plexilandia/graphs");
//se cierra el modelo.
model.closeModel();
}
}
```

En el ejemplo anterior se utilizó gran parte de los hostpots por defecto, es decir, valores por defecto. Sin embargo, el framework provee interfaces que permiten al desarrollador especificar la función de distancia para la compración de vectores y la función de cálculo de pesos para las aristas de los grafos SNA.

3.2.6. Publicación en SourceForge

Uno de los objetivos específicos es la publicación del framework en algún sitio web. El portal escogido fue SourceForge[6]. El proyecto se encuentra en:

- http://sourceforge.net/projects/siocextended/

Cabe señalar que el proyecto fue subscrito bajo la licencia GNU General Public License[7] (GPL), es decir, el framework es un software libre y nadie puede restringir el acceso a éste. Al momento de la entrega de éste informe, el proyecto se encontraba en estado alfa, sin embargo, una vez que se carguen todos los ejemplo y tutoriales, se cambiará a un estado beta, al menos.

[6]http://sourceforge.net [Fecha último acceso: marzo de 2011]
[7]http://www.gnu.org/licenses/licenses.es.html [Fecha último acceso: marzo de 2011]

Capítulo 4

Validación de la Solución

Para la validación del framework se desarrolló una aplicación, en base a los componentes del framework, para para llevar el foro de la comunidad virtual de Plaxilandia a la Web Semántica. Esta comunidad cuenta con más de 2500 miembros desde marzo del 2002, sin embargo, el periodo procesado corresponde al año 2009, en el cual se registraron 440 usuarios activos, los que generaron 10.546 posts. La aplicación desarrollada siguió los pasos expuestos en la sección 3.2.5.

Dada la gran cantidad de posts en estudio (10.546), no es posible mostrar todo el RDF en este informe, sin embargo, el Anexo D muestra un extracto del RDF (en N3) generado a partir del foro de Plexilandia.

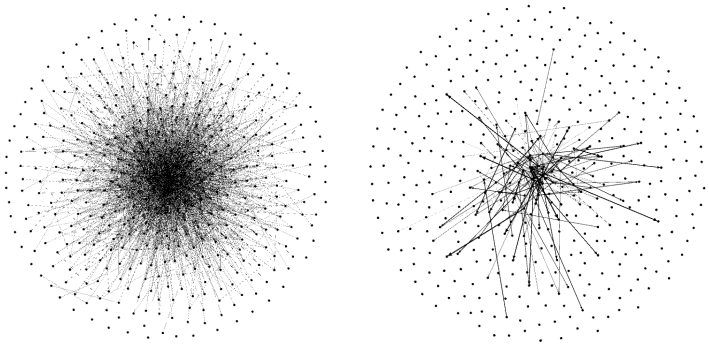
Luego, Utilizando el exportador de redes para SNA, se generó las topologías del creador, con y sin LDA. Cabe mencionar que los vectores de LDA estaban pre-calculados. Para el caso del grafo con LDA, se filtraron las aristas con peso menor a 0,8 y como medida de similitud entre vectores se utilizó el coseno. Todos lo grafos fueron exportados a archivos Pajek.

Luego para el análisis de redes sociales, se utilizó la herramienta Gephi[1]. Gephi es un framework diseñado para el análisis de redes sociales mediante la visualización y el cálculo de indicadores estándares. El framework desarrollado puede potenciarse con el uso de una herramienta como Gephi, ya que el framework permite exportar los foros (representados semánticamente) a archivos Pajek y Gephi puede importar esos archivos para aplicar análisis de redes de sociales.

[1] http://gephi.org/ [Fecha último acceso: marzo de 2011]

La Figura 4.1 muestra las redes de la comunidad virtual Plexilandia, en base a la topología del creador, con y sin LDA. De la figura se puede desprender que la red en base a LDA es mucho menos densa. La red del creador-LDA tiene una desidad de 0,001, en contra de la red del creador cuya densidad es 0,013. La diferencia de densidad se explica porque la red del creador tiene 3094 aristas, mientras que la red del creador con LDA sólo tiene 200 (7 % de la red del creador).

Además, en el caso de la red del creador-LDA, se puede desprender que las aristas son de distinto espesor, en cambio en la red sin LDA, todas sus aristas son iguales. Esto se debe a que los pesos de las aristas de la redes con LDA son cálculados mediante la distancia entre los vectores de conceptos, en cambio en la redes sin LDA, los pesos son por defecto igual a 1.0.



(a) Creador (b) Creador-LDA

Figura 4.1: Visualización red Plexilandia

Luego se aplicó el algorimo de Hits a las topologías en estudio. Hits se base en dos principios: Si exiten muchas recomendaciones (aristas) hacia un nodo n, entoces ese nodo es importante; y si los recomendadores son importantes, entoces el nodo n es aún más importante. Hits provee de dos indicadores: authority y hub. Un nodo es buen authority si es un buen recomendador (muchas aristas salientes) y es buen hub si tiene muchas recomendaciones (muchas aristas entrantes).

Los resultados de aplicar Hits a los grafos SNA se muestran en Tabla 4.1 (Mejores 10). Se puede desprender de la tabla que existen tres subconjuntos[2]: un grupo de usuarios que se encuentra sólo en la columna del creador, un grupo que se encuentra sólo en la columna del creador-LDA y un grupo que se encuentra en ambas columnas (marcados con negrita).

El grupo que se encuentra en ambas columnas, es un grupo interesante, ya que sus miembros no sólo generan gran cantidad de contenido y participación en la comunidad, sino

[2]Nota: Los nombres de usuarios reales fueron omitidos

que también generan rico contenido para el resto de los miembros.

El grupo que se encuentran sólo en la columna del creador, puede corresponder a miembros que generan mucho contenido y alta participación, pero no necesariamente rico contenido. Y por el contrario, el grupo que se encuentra sólo en la columna del creador, es aquel que genera rico contenido y alta participación en torno a éste. LDA pondera mucho el significado de las interacciones entre los miembros de un foro, por eso los usuarios que generan mayor contenido no necesariamente tienen buen ranking con LDA.

	Creador		Creador -LDA	
	Autorithy	Hub	Autority	Hub
1	**Usuario1**	**Usuario1**	**Usuario7**	**Usuario7**
2	Usuario2	Usuario2	**Usuario3**	**Usuario1**
3	Usuario3	**Usuario4**	**Usuario1**	**Usuario3**
4	**Usuario4**	Usuario10	Usuario11	Usuario11
5	Usuario5	Usuario3	Usuario2	Usuario13
6	Usuario6	Usuario6	Usuario12	Usuario17
7	**Usuario7**	Usuario5	Usuario13	Usuario14
8	Usuario8	Usuario9	Usuario14	Usuario18
9	Usuario9	**Usuario7**	Usuario15	Usuario15
10	Usuario10	Usuario8	Usuario16	Usuario12

Cuadro 4.1: Ranking usuarios según Hits en base a la topología del creador, con y sin LDA

En esta sección se mostró, a modo de ejemplo, como realizar análisis de redes sociales a los foros web semánticos utilizando el framework y Gephi. De esta misma firma, cada desarrollador o investigador podría aplicar otros algoritmos a las redes, con el fin de obtener información relevante. Además, se podrían construir redes en base a la información conjunta de más de un sitio, aprovechando las ventajas de la Web Semántica.

Capítulo 5

Conclusiones

A pesar del gran éxito que han tenido los sitios web sociales, están aislados entre sí como islas en el mar. No es posible reutilizar la información que estos generan. Si un usuario desea tener acceso a dos sitios diferentes, debe registrarse en cada uno de ellos, ya que no es posible reutilizar el perfil. Los sitios web sociales no fueron construidos para colaborar entre ellos.

Sin embargo el problema se origina en la creación de la Web. La Web fue creada para compartir documentos HTML y no datos y las máquinas no pueden interpretar documentos HTML. Como respuesta a esta necesidad surgió la Web Semántica.

La Web Semántica propone crear un ambiente para que los datos o la información publicada en la Web sean interpretables por humanos y por máquinas. En ese sentido la Web Semántica puede ser la solución para interconectar lo sitios web sociales.

Como parte de la solución a este problema, en este trabajo se propuso el desarrollo de un framework que permita llevar sitios web sociales a la Web Semántica, el cual se llevó a cabo con éxito. El principal objetivo, de llevar sitios web sociales a la Web Semántica fue logrado. Además se desarrollaron herramientas para que fuese posible aplicar análisis de redes sociales sobre la estructura semántica generada por el mismo framework. Luego, utilizando programas o herramientas para el análisis de redes sociales, como Gephi, es posible aplicar análisis de redes sociales a los grafos o redes generadas por el framework.

Mediante el caso de estudio: Plexilandia, fue posible mostrar el comportamiento y el uso del framework en una comunidad real. Sobre el cual y a modo de ejemplo, se hizo un

pequeño análisis de redes sociales.

Este trabajo representa una contribución a la Web Semántica y al análisis de redes sociales. Ambas áreas se ven directamente beneficiadas con el desarrollo de este trabajo. Sin embargo, este framework no pretende ser una solución final, sino entregar un conjunto de herramientas iniciales que permita a los investigadores de la Web facilitar su trabajo y un punto de partida para seguir trabajando en la mejora de la Web actual.

5.1. Trabajo Futuro

Unos de los puntos que el framework no consideró, fue el desarrollo de técnicas, como LDA y Concept-Based en base la representación semántica provista por éste. En ese sentido, éste debiese ser unos de los primeros trabajos a abordar en el futuro.

Otro trabajo importante, es seguir extendiendo la ontología de SIOC para que permita representar semánticamente las interacciones sociales entre los miembros de un sitio web, u otros datos que permitan crear innovadores servicios para los miembros de los sitios web sociales.

El framework tampoco contempló la generación de indicadores de SNA, sólo se limitó a extraer los grafos de interacciones. Sin embargo, puede ser útil contar con algunos indicadores de redes sociales a priori, es decir, al momento de la generación (exportación) de los grafos SNA.

5.2. Publicaciones

El desarrollo de este trabajo sirvió de apoyo para la publicación del paper "Leveraging Social Network Analysis with Topic Models and the Semantic Web" en la WI[1]. En dicho paper se propuso mejorar el análisis de redes sociales mediante el uso de la Web Semántica. Para ello se hizo una extensión a la ontología SIOC, más amplia a la expuesta en este informe, que pretendía expresar explícitamente las relaciones o interacciones entre los miembros de una comunidad.

Luego, a partir de los datos semánticos, se generaron los grafos para SNA (con y sin

[1]http://www.wi-consortium.org/ [Fecha último acceso: marzo de 2011]

LDA), sobre los cuales se aplicó Hits para obtener los miembros claves. Al igual que en este trabajo, en los papers se utilizó la comunidad de Plexilandia como objeto de estudio.

En el desarrollo del paper, se trabajó en conjunto con Tope Omitola[2], Sebastián Ríos [3] y Felipe Aguilera[4].

[2]Intelligence, Agents, Multimedia (IAM) Group School of Electronics and Computer Science, University of Southampton, UK
[3]Departamento Ingeniería Industrial, Universidad de Chile
[4]Departamento Ciencias de la Computación, Universidad de Chile

Referencias

[1] S. Allen, S. Evans, y D. Ure, "Virtual communities of practice: vehicles for organisational learning and improved job performance," *Int. J. Learn. Technol.*, vol. 1, no. 3, pp. 252–272, 2005.

[2] H. Alvarez, S. A. Ríos, F. Aguilera, E. Merlo, y L. Guerrero, "Enhancing social network analysis with a concept-based text mining approach to discover key members on a virtual community of practice," in *Proceedings of the 14th international conference on Knowledge-based and intelligent information and engineering systems: Part II*, ser. KES'10. Berlin, Heidelberg: Springer-Verlag, 2010, p. 591–600.

[3] T. Berners-Lee, J. Hendler, y O. Lassila, "The semantic web," *Scientific American Magazine*, May 2001.

[4] U. Bojārs, J. G. Breslin, V. Peristeras, G. Tummarello, y S. Decker, "Interlinking the social web with semantics," *IEEE Intelligent Systems*, vol. 23, no. 3, pp. 29–40, 2008.

[5] J. Bosch, M. Molin, M. Mattson, y P. Bengtsson, "Building application frameworks, chapter object-oriented frameworks - problems & experiences." *Wiley and Sons*, 1999.

[6] J. G. Breslin, A. Passant, y S. Decker, *The Social Semantic Web*, 1st ed. Springer, Oct. 2009.

[7] S. Brin y L. Page, "The anatomy of a large-scale hypertextual web search engine," in *Computer Networks and ISDN Systems*, vol. 30. Amsterdam, The Netherlands, The Netherlands: Elsevier Science Publishers B. V., Abr. 1998, p. 107–117.

[8] H. Ellonen, M. Kosonen, y K. Henttonen, "The development of a sense of virtual community," *Int. J. Web Based Communities*, vol. 3, no. 1, pp. 114–130, 2007.

[9] E. Gamma, R. Helm, R. Johnson, y J. M. Vlissides, *Design Patterns: Elements of Reusable Object-Oriented Software*, 1st ed. Addison-Wesley Professional, Nov. 1994.

[10] S. Gerd, A. Hotho, y B. Berendt, "Semantic web mining," *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 4, p. 124–143, Jun. 2006.

[11] J. Hanisch y D. Churchman, "Virtual communities of practice: the communication of knowledge across cultural boundaries," *Int. J. Web Based Communities*, vol. 4, no. 4, pp. 418–433, 2008.

[12] P. Isaías, P. Miranda, y S. Pífano, "Critical success factors for web 2.0 — a reference framework," in *Proceedings of the 3d International Conference on Online Communities and Social Computing: Held as Part of HCI International 2009*, ser. OCSC '09. Berlin, Heidelberg: Springer-Verlag, 2009, p. 354–363.

[13] R. Johnson, "How to design frameworks." *Tutorial Notes for the 1993 Conference on Object Oriented Programming, Systems, Languages and Systems (OOPLSA 1993)*, 1993.

[14] R. E. Johnson y B. Foote, "Designing reusable classes," *Journal of Object-Oriented Programming (1988)*, vol. 1, no. 2, pp. 22–35, 1988.

[15] W. Kim, O. Jeong, y S. Lee, "On social web sites," *Information Systems*, vol. 35, p. 215–236, Abr. 2010.

[16] S. Kinsella, J. G. Breslin, A. Passant, y S. Decker, "Applications of semantic web methodologies and techniques to social networks and social websites," *Reasoning Web*, p. 171–199, 2008.

[17] J. M. Kleinberg, "Authoritative sources in a hyperlinked environment," in *Proceedings of the ninth annual ACM-SIAM symposium on Discrete algorithms*, ser. SODA '98. Philadelphia, PA, USA: Society for Industrial and Applied Mathematics, 1998, p. 668–677.

[18] M. Kosonen, "Knowledge sharing in virtual communities; a review of the empirical research," *Int. J. Web Based Communities*, vol. 5, no. 2, pp. 144–163, 2009.

[19] G. L'Huillier, S. A. Ríos, H. Alvarez, y F. Aguilera, "Topic-based social network analysis for virtual communities of interests in the dark web," in *ACM SIGKDD Workshop on Intelligence and Security Informatics*, ser. ISI-KDD '10. New York, NY, USA: ACM, 2010, p. 9:1–9:9.

[20] H. Mili, F. Mili, y A. Mili, "Reusing software: Issues and research directions," *IEEE Trans. Softw. Eng.*, vol. 21, no. 6, pp. 528–562, 1995.

[21] P. Penfold, "Virtual communities of practice: Collaborative learning and knowledge management," in *Proceedings of the 2010 Third International Conference on Knowledge Discovery and Data Mining*. IEEE Computer Society, 2010, pp. 482–485.

[22] J. T. Pollock, *Semantic Web For Dummies*. Wiley, Mar. 2009.

[23] S. A. Ríos, F. Aguilera, y L. A. Guerrero, "Virtual communities of practice's purpose evolution analysis using a Concept-Based mining approach," in *Proceedings of the 13th International Conference on Knowledge-Based and Intelligent Information and Engineering Systems: Part II*. Santiago, Chile: Springer-Verlag, 2009, pp. 480–489.

[24] D. Roberts y R. Johnson, "Evolving frameworks: A pattern language for developing Object-Oriented frameworks," *Proceedings of the Third Conference on Pattern Languages and Programming*, vol. 3, 1996.

[25] J. P. Scott, *Social Network Analysis: A Handbook*, 2nd ed. Sage Publications Ltd, Mar. 2000.

[26] J. van Gurp y J. Bosch, "Design, implementation and evolution of object oriented frameworks: concepts and guidelines," *Softw. Pract. Exper.*, vol. 31, no. 3, pp. 277–300, 2001.

[27] I. Varlamis y I. Apostolakis, "Self-supportive virtual communities," *Int. J. Web Based Communities*, vol. 6, no. 1, pp. 43–61, 2010.

[28] M. Vásquez, "Desarrollo de un framework para el problema de ruteo de vehículos," Tésis para optar al grado de Magíster en Gestión de Operaciones, Departamento de Ingeniería Industrial, Facultad de Ciencias Físicas y Matemáticas. Universidad de Chile, Chile, 2007.

[29] J. D. Velásquez y L. C. Jain, "Web intelligence on the social web," in *Advanced Techniques in Web Intelligence -1*, 1st ed. Springer, Sep. 2010.

[30] B. Wellman, "Changing connectivity: A future history of Y2.03K," *Sociological Research Online*, vol. 4, no. 4, Feb. 2000.

[31] B. Wellman, "For a social network analysis of computer networks: a sociological pers-

pective on collaborative work and virtual community," in *Proceedings of the 1996 ACM SIGCPR/SIGMIS conference on Computer personnel research*. Denver, Colorado, United States: ACM, 1996, pp. 1–11.

[32] N. Zhong, J. Liu, y Y. Yao, *Web intelligence*. Springer, Abr. 2003.

ANEXOS

Anexo A

Configuración SDB en Windows

1. Se require tener instalado MySQL (5.1 o más reciente).
2. Crear base de datos para SDB
3. Bajar e instalar el driver jdbc para mysql en C:\extra
   a) Obtener la versión 5.1.8 de http://dev.mysql.com/downloads/connector/j/5.1.html
   b) Descargar y extraer en C:\extra (extraer en C:\extra\mysql-connector-java-5.1.8).
4. Descargar SDB 1.3 de http://jena.hpl.hp.com/wiki/SDB e instalar en c:\extra (extraer en C:\extra\SDB-1.3.0)
   a) Copiar C:\extra\SDB-1.30\Store\sdb_mysql_innodb.ttl al archivo C:\extra\SDB-1.30\sdb.ttl
5. Instalar cygwin
   a) Obtenerlo desde http://cygwin.com/
      1) Instalar con la configuración por defecto, excepto agregar el paquete vim (En la edición de categorías)
      2) Installar en C:/cygwin
6. Ejecutar cygwin y configurar la variable de entorno.
   a) Crear un directorio para un usuario: mkdir /<tunombre>
      1) En Windows Inicio->Panel de Control->Sistema -> Avanzado -> Variables de Entorno.
      2) En las variables de usuario agregar

a' Nombre de variable: HOME

b' Valor: C:\cygwin\<tunombre>

3) Reiniciar cygwin

b) Crear las variables de entorno necesarias para SDB

1) Escribir las siguientes lineas en el archivo .bash_login:

```
# (por http://jena.hpl.hp.com/wiki/SDB/Commands)
export SDBROOT="c:/extra/SDB-1.3.0"
export SDB_USER="el usuario de la base de datos"
export SDB_PASSWORD="la contraseña del usuario"
export SDB_JDBC=
"/extra/mysql-connector-java-5.1.8/mysql-connector-java-5.1.8-bin
    .jar"
PATH=$SDBROOT/bin:$PATH
```

2) En cygwin:

a' d2u .bash_login

b' source .bash_login

7. Probar la configuración SDB (por http://jena.hpl.hp.com/wiki/SDB/Installation)

a) cd C:/extra/SDB-1.3.0

b) bin/sdbconfig –sdb=sdb.ttl –create

1) No deberían haber alertas.

c) bin/sdbtest –sdb=sdb.ttl testing/manifest-sdb.ttl

1) Debería salir un mensaje en la pantalla, la última linea debería ser algo así: Tests = 82 : Successes = 82 : Errors = 0 : Failures = 0

8. SDB está listo para ser usados por Jena.

Anexo B

Extracto de código del exportador de PHPBB2 a SIOC

```
//Se instancia EagerSite
EagerSite site = new EagerSite(http://plexilandia.cl/foro,
   Plexilandia , null,
http://plexilandia);
// se obtienen todas las categorías
ResultSet rsCat = conn.getCategories();
if (rsCat != null) {
    while (rsCat.next()) {
    // se crean las categorías
    EagerForum cat = new EagerForum(catId, catUrl, catTitle),
       null);
    // se obtienen todos los foros
    ResultSet rsForum = conn.getForumByCatID(catId);
    if (rsForum != null) {
        while (rsForum.next()) {
        // se crean los foros
        EagerForum forum =
        new EagerForum(fId, fUrl, forumName, forumDesc);
        // se obtienen todos los tópicos
        ResultSet rsTopics =
        conn.getTopicsByForumID(rsForum.getString("forum_id"));
        if (rsTopics != null) {
            while (rsTopics.next()) {
            // se crean los tópicos (posts)
            EagerPost topic = new EagerPost(tId, tUrl,
               tPostSubject,
            tPostText, "utf8", tCreator, tPostTime, tPostEditTime
               );
            // se obtienen las respuestas
            ResultSet rsReplies = conn.getRepliesByTopicID(tId);
            // esto es para distinguir las respuestas
```

```
            int replyNum = 1;
            while (rsReplies.next()) {
            // se crean las respuestas
            EagerPost reply = new EagerPost(rId, rUrl,
               rPostSubject,
            rPostText, "utf8", rCreator, rPostTime, rPostEditTime
               );
            // se agregan las respuestas a los tópicos
            topic.addReply(reply);
        }
        // se agregan las tópicos a los foros
        forum.addTopic(topic);
        }
    }
    // se agregan los foros a las categorías
    cat.addForum(forum);
    }
    }
// se agregan las categorías a los foros
site.addForum(cat);
}
}
    // se obtienen los usuarios
ResultSet rsUsers = conn.getUsers();
if (rsUsers != null) {
    while (rsUsers.next()) {
    // se crean los usuarios
    EagerUser u = new EagerUser(uId, uUrl , username, email);
    // se agregan los usuarios al sitio
site.addUser(u);
}
}
```

Anexo C

Ejemplo de archivo de configuración para el exportador de PHPBB2

```
# Datos Conexion MYSQL.
servidor = localhost
puerto = 3306
usuario=root
clave=root
base=plexilandia
# Nombres de tablas
CATEGORIES_TABLE=plxcl_phpbb_categories
FORUMS_TABLE=plxcl_phpbb_forums
TOPICS_TABLE = plxcl_phpbb_topics
USERS_TABLE = plxcl_phpbb_users
POSTS_TABLE = plxcl_phpbb_posts
POSTS_TEXT_TABLE = plxcl_phpbb_posts_text
FAV_TERMINOS_TABLE = fav_terminos
FAV_LDA_PUNTAJE_TABLE = fav_lda_puntaje
FAV_LDA_CONCEPTOS_TABLE= fav_lda_conceptos
#Metadatos del sitio
homepage = http://www.plexilandia.cl
language = es
name = Plexilandia description = este es un foro de plex
```

Anexo D

Extracto de Plexilandia en RDF-N3

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>.
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>.
@prefix foaf: <http://xmlns.com/foaf/0.1/>.
@prefix dcterms: <http://purl.org/dc/terms/>.
@prefix sioc: <http://rdfs.org/sioc/ns#>.
@prefix xsd: <http://www.w3.org/2001/XMLSchema#>.

<www.plexilandia.cl/Site1> rdf:type     sioc:Site;
        dc:description  "este es un foro de plex."@es;
        sioc:topic      <http://www.dmoz.org/Computers/Internet/
           On_the_Web/Message_Boards/>;
        foaf:homepage   <http://www.plexilandia.cl>;
        sioc:host_of <www.plexilandia.cl/Site1/Plexilandia>.
                <www.plexilandia.cl/Site1/Plexilandia> rdf:type
                   sioc:Forum;
        sioc:parent_of
                <www.plexilandia.cl/Site1/Plexilandia/
                   Amplificadores>.

<www.plexilandia.cl/Site1/Plexilandia/Amplificadores> rdf:type
   sioc:Forum;
        dc:description "todo sobre amplificadores, tubos, etc."
           @es;
        sioc:container_of               <www.plexilandia.cl/Site1
           /Plexilandia/Amplificadores/Post1>;    sioc:
           container_of               <www.plexilandia.cl/Site1/
           Plexilandia/Amplificadores/Post2>.

<www.plexilandia.cl/Site1/Plexilandia/Amplificadores/Post1> rdf:
   type sioc:Post;
        dc:title        "Presentación y elección de amplficador"
           @es ;
```

```
        dcterms:created          "1230906979"^^xsd:date ;
        dcterms:modified         "null"^^xsd:string ;
        sioc:content             "Primero que nada ...."^^xsd:
           string ;
        content:encoded          "..."^^xsd:string ;
        sioc:has_creator         <http://www.username1> ;
        sioc:has_reply
                <www.plexilandia.cl/Site1/Plexilandia/
                   Amplificadores/Post1/Reply1>,
                <www.plexilandia.cl/Site1/Plexilandia/
                   Amplificadores/Post1/Reply2>.

<www.plexilandia.cl/Site1/Plexilandia/Amplificadores/Post1/Reply1
   > rdf:type sioc:Post;
        dcterms:created          "1230949014"^^xsd:date;
        dcterms:modified         "null"^^xsd:string ;
        sioc:content             "Bueno muchas ...."^^xsd:string ;
        content:encoded          "..."^^xsd:string ;
        sioc:has_creator         <http://www.username2>.

<http://www.username1> rdf:type sioc:UserAccount;
        foaf:homepage <http://www.username1.com> ;
        foaf:nick "bustos"^^xsd:string ;
```