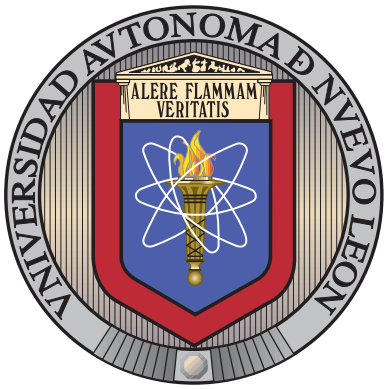
Universidad Autónoma de Nuevo León

Facultad de Ingeniería Mecánica y Eléctrica

División de Estudios de Posgrado



Construcción de perfiles de usuario a partir de repositorios de información personal

por

Rafael Olivares Arredondo

en opción al grado de

Maestría en Ingeniería de la Información

con Orientación en Informática

San Nicolás de los Garza, Nuevo León marzo 2013

Universidad Autónoma de Nuevo León

Facultad de Ingeniería Mecánica y Eléctrica

División de Estudios de Posgrado

Construcción de perfiles de usuario a partir de repositorios de información personal

por

Rafael Olivares Arredondo

en opción al grado de

Maestría en Ingeniería de la Información

con Orientación en Informática

San Nicolás de los Garza, Nuevo León marzo 2013

Universidad Autónoma de Nuevo León

Facultad de Ingeniería Mecánica y Eléctrica

División de Estudios de Posgrado

Los miembros del Comité de Tesis recomendamos que la Tesis «Construcción de perfiles de usuario a partir de repositorios de información personal», realizada por el alumno Rafael Olivares Arredondo, con número de matrícula 0842446, sea aceptada para su defensa como opción al grado de Maestría en Ingeniería de la Información con Orientación en Informática.

El Comité de Tesis

Dra. Sara Elena Garza Villarreal
Asesor

Dr. Francisco Torres Guerrero
Revisor

M.C. Aída Lucina González Lara
Revisor

Vo. Bo.

Dr. Moisés Hinojosa Rivera
División de Estudios de Posgrado

San Nicolás de los Garza, Nuevo León, marzo 2013

Este trabajo esta dedicado

Al recuerdo de mis abuelos

A la fortaleza de mis padres

Al amor de mis hermanas

A la llegada de Andrea

A la energía de Paty

A la mirada de Santiago

Al abrazo de mis tíos Paty y Pancho

Índice general

Agradecimientos xiv

Resumen xv

1. Introducción 1

1.1. Motivación y justificación 2

1.2. Definición del problema 3

1.3. Protocolo de investigación 4

1.3.1. Objetivos 4

1.3.2. Preguntas de investigación 5

1.3.3. Hipótesis 5

2. Antecedentes 6

2.1. Perfiles 7

2.2. Sistemas de recomendación 9

2.3. Minería de texto 13

2.3.1. Minería de texto y recuperación de información 15

2.3.2. Modelo Espacio Vectorial 16

2.3.3. Similitud Cosenoidal 17

2.4. Trabajos relacionados 18

2.5. Resumen 20

3. El Modelo **21**

3.1. Marco formal de trabajo 21

3.1.1. Procesamiento del repositorio 23

3.2. Generación del perfil 24

3.3. Visualización del perfil 27

3.4. Enriquecimiento del perfil 31

3.5. Resumen 34

4. Caso de estudio - Marcadores de Internet **35**

4.1. Los marcadores de Internet como fuente de información implícita . . 35

4.2. Obtención del repositorio 37

4.2.1. Obtención de las páginas web a partir de los marcadores . . . 37

4.2.2. Extracción de texto 40

4.2.3. Limpieza 45

4.3. Resumen 57

5. Experimentación, pruebas y resultados **58**

5.1. Creación de perfiles de usuario 59

5.1.1. Muestra de usuarios y repositorios 59

5.1.2. Perfiles de usuario 62

5.2. Calificación de perfiles de usuario 62

5.2.1. Configuración 63

5.2.2. Resultados 64

5.2.3. Discusión de resultados 67

5.3. Recuperación de grupos de usuarios 70

5.3.1. Matrices de similitud 71

5.3.2. Configuración 71

5.3.3. Resultados 73

5.3.4. Discusión de resultados 81

5.3.5. Discusión 82

5.4. Discusión general y resumen del capítulo 82

6. Conclusiones y trabajos futuros **85**

6.1. Resumen de lo propuesto en la tesis 85

6.2. Respuestas a las preguntas de investigación 86

6.3. Contribuciones 86

6.4. Comentarios conclusivos 87

6.5. Trabajos futuros 87

A. Palabras vacías (StopWords) **89**

B. Evaluaciones de los usuarios **92**

C. Cohesión y Separación **93**

Índice de figuras

3.1. Ejemplo de Nube de etiquetas. 28

3.2. Nube de palabras procesadas mediante TFIDF. 28

3.3. Nube de palabras procesadas mediante TFIDF con repositorio pequeño. 30

3.4. Nube de palabras procesadas mediante TFIDF y enriquecida con WordNet. 33

4.1. Algunas fuentes de información del usuario. 36

4.2. Muestra parcial de un documento de exportación de marcadores. . . . 38

4.3. Estructura de HTML básica para generar una página web. 40

4.4. Caracteres a sustituir. 48

4.5. Signos de puntuación eliminados. 49

5.1. Imagen *Dummy* mostrada a los usuarios en la evaluación. 62

5.2. Imagen con estrellas para evaluar por los usuarios. 64

5.3. Nube de palabras procesadas mediante TFIDF. 68

5.4. Nube de palabras procesadas mediante TFIDF y enriquecida con WordNet. 68

5.5. Nube de palabras generada con TFIDF de repositorio pequeño. . . . 70

5.6. Nube de frecuencia de palabras de repositorio pequeño. 70

5.7. Muestra de caso ideal de cohesión. 72

5.8. TFIDF colección completa, 2 grupos. 74

5.9. Enriquecidos colección completa, 2 grupos. 74

5.10. TFIDF colección reducida, 2 grupos. 75

5.11. Enriquecidos colección reducida, 2 grupos. 75

5.12. TFIDF colección completa, 3 grupos. 76

5.13. Enriquecidos colección completa, 3 grupos. 76

5.14. TFIDF colección reducida, 3 grupos. 77

5.15. Enriquecidos colección reducida, 3 grupos. 77

5.16. Resultados de relación entre usuarios de perfiles enriquecidos. 79

5.17. Resultados de relación entre usuarios de perfiles generados con TFIDF. 80

Índice de tablas

2.1. Representación matricial del modelo espacio vectorial 16

3.1. Ejemplo de una visualización que incluye atributos para color (RGB), posición en 2D (x,y) y tamaño de letra. 23

3.2. Muestra parcial de un documento previo al cálculo de TF. 26

3.3. Resultados TF de documento procesado. 26

3.4. Documento procesado con TFIDF. 27

3.5. Algunas relaciones con las que WordNet asocia synsets. 32

4.1. Etiquetas de HTML consideradas para la extracción de contenido. . . 42

4.2. Ejemplo parcial de contenido del texto generado a través de la aplicación. 44

4.3. Números utilizados para sustituir letras. 47

4.4. Ejemplo parcial de contenido del texto ya procesado por el primer mecanismo de limpieza. 49

4.5. Lista breve de palabras vacías. 50

4.6. Enlaces encontrados en CMS para interactuar con el contenido. . . . 52

4.7. Enlaces que los CMS manejan como base en sus plantillas. 53

4.8. Enlaces que los CMS manejan en menús y/o sub menús. 54

4.9. Enlaces que los CMS manejan referente al registro de usuarios. 55

4.10. Enlaces que los CMS manejan para tener un acercamiento con los usuarios. 55

4.11. Enlaces que los CMS manejan para la interacción con el usuario. . . . 56

4.12. Lista de enlaces no clasificados. 56

5.1. Representación de la muestra. 59

5.2. Tamaño de los repositorios de los usuarios. 60

5.3. Estadística descriptiva de la muestra utilizada. 61

5.4. Grupo de usuarios con mayor número de documentos a procesar. . . . 61

5.5. Grupo de usuarios con menor número de documentos a procesar. . . . 61

5.6. Valor promedio extraído de los resultados de las evaluaciones. 65

5.7. Resultados del grupo de usuarios con mayor número de documentos a procesar. 66

5.8. Valor promedio del grupo de usuarios con mayor número de documentos. 66

5.9. Resultados del grupo de usuarios con menor número de documento. . 67

5.10. Valor promedio del grupo de usuarios con menor número de documento. 67

5.11. Valor de la relación entre los usuarios licenciatura. 78

5.12. Valor de la relación entre los usuarios maestría. 78

5.13. Valores promedio de cohesión y separación. 81

5.14. Valor de la relación entre los usuarios fuera de la muestra estudiantil. 81

5.15. Comparativa de valores promedio de los grupo de usuario con menor y mayor número de documentos. 83

B.1. Resultados de las evaluaciones de las visualizaciones de los perfiles de los usuarios. 92

C.1. TFIDF colección completa, 2 grupos 93

C.2. Enriquecidos colección completa, 2 grupos 93

C.3. TFIDF colección completa, 3 grupos 93

C.4. Enriquecido colección completa, 3 grupos 94

C.5. TFIDF colección reducida, 2 grupos 94

C.6. Enriquecidos colección reducida, 2 grupos 94

C.7. TFIDF colección reducida, 3 grupos 94

C.8. Enriquecido colección reducida, 3 grupos 94

Agradecimientos

A la Dra. Sara Elena Garza Villarreal por su paciencia, dedicación y vocación a la ciencia, sin su ayuda este trabajo no hubiera sido posible.

A mis revisores de tesis, la M.C. Aída Lucina González Lara y el Dr. Francisco Torres Guerrero quienes dedicaron tiempo y esfuerzo en la revisión de tesis y me apoyaron en cada etapa como estudiante con sus consejo.

A la Dra. Satu Elisa Schaeffer por haber sembrado en mi la primera semilla que inició todo el proceso, por su tiempo y su vocación a la enseñanza.

A mi hermano Fer, por su apoyo y por nunca dejar de crear y soñar.

A todos aquellos que se siguen burlando cada vez que les digo que voy hacer un doctorado :)

Y por último pero no menos importante; a la Universidad Autónoma de Nuevo León y a mi casa la Facultad de Ingeniería Mecánica y Eléctrica por estar en la búsqueda constante de la excelencia académica.

Resumen

Rafael Olivares Arredondo.

Candidato para el grado de Maestro en Ingeniería de la Información
con Orientación en Informática.

Universidad Autónoma de Nuevo León.
Facultad de Ingeniería Mecánica y Eléctrica.

Título del estudio:

Construcción de perfiles de usuario a partir de repositorios de información personal

Número de páginas: 100.

Objetivos y método de estudio: Durante la investigación desarrollada en esta tesis, se realiza un proceso de obtención de perfiles de usuario mediante repositorios de información personal. El perfil del usuario es conformado por intereses implícitos encontrados a través de técnicas de minería de texto, y puede ser enriquecido con fuentes externas de información. Para generar el perfil, se lleva a cabo un proceso de extracción de texto y de limpieza del repositorio con el fin de conformar un *saco de palabras* por documento, del cual más tarde se seleccionan— mediante **TFIDF** (*Text Frequency Inverse Document Frequency*)— las palabras más significativas. El perfil del usuario consiste en la concatenación de estos documentos. Debido a que algunos

perfiles contienen poca información, se propuso un mecanismo de enriquecimiento de los perfiles a través de una entidad semántica (**Wordnet**) que agrega palabras relacionadas con las palabra del perfil. Utilizamos también una visualización de nube de palabras (*tag cloud*) para representar el perfil de usuario, con los perfiles conformados, utilizamos la *similitud cosenoidal* para sacar la relación que existe entre los perfiles de los usuarios e identificar grupos de usuarios con intereses similares.

Para validar el método, se obtuvieron los repositorios de una muestra de usuarios, se generaron sus perfiles y se generaron las respectivas visualizaciones. Estas visualizaciones fueron evaluadas por los usuarios con pleno desconocimiento de que eran sus perfiles y se obtuvieron resultados favorables. Así mismo, con los perfiles conformados, utilizamos *similitud cosenoidal* para obtener la relación entre los usuarios y recuperar grupos con intereses similares.

CONTRIBUCIONES Y CONCLUSIONES: En el preprocesamiento de los repositorios, se propuso generar una lista de palabras que son agregadas por los manejadores de contenido (*CMS, Content Management System*) para ser discriminadas. Esto debido a que si no son eliminadas, salen en los perfiles como temas de interés de los usuarios. Esto impacta sobre todo en repositorios que son pequeños. Sobre el mismo tema de repositorios pequeños o escasos de información, contribuimos con un mecanismo de enriquecimiento de perfiles a través de una entidad semántica la cual contribuye agregando palabras relacionadas a las que contiene el perfil.

Se aportó un modelo conceptual y marco formal de trabajo de donde se genera un perfil de usuario de un repositorio de información personal; también se expuso una visualización de *nube de palabras* para los perfiles.

Los resultados de la investigación muestran que los usuarios se sintieron identificados con las visualizaciones de los perfiles y encontramos que a través de la información implícita que se encuentra en repositorios de información personal, podemos relacionar usuarios encontrando los intereses que tienen en común. Concluimos que los resultados de la investigación nos muestran que debemos seguir trabajando en

experimentos para encontrar cifras más acertadas. Es necesario probar más mecanismos que puedan enriquecer el proceso de perfiles y explorar los nuevos escenarios a los que los usuarios se enfrentan hoy en día como los dispositivos móviles y la actividad de redes sociales.

Firma del asesor: _____

Dra. Sara Elena Garza Villarreal

Capítulo 1

Introducción

El estudio titulado *The Digital Universe Decade - Are You Ready?* realizado en el mes de mayo de 2010 por la compañía **IDC iView**, muestra que en el 2009 se estableció una nueva marca en la cantidad de información digital creada y replicada (a la cual se le conoce como el *universo digital*). Se estima que los 800,000 petabytes[1] de ese año aumenten a una cifra 44 veces mayor es decir, de 32 zettabytes[2], en el 2020. Tales cantidades de información, según este estudio, crean un estado de estrés (*infoxicación* Cornella (2000)) entre los usuarios para administrarla, almacenarla, protegerla y poder disponer de ella. Al mismo tiempo, generan diversos cuestionamientos relacionados con áreas como el almacenamiento y acceso de la información; por ejemplo: *¿Cómo vamos a encontrar la información que necesitemos cuando la necesitemos?* y *¿Cómo vamos a saber qué información necesitamos guardar y cómo la vamos a guardar?*. En el análisis realizado por **IDC iView** de este panorama tan complejo, concluyen que serán necesarias nuevas herramientas y técnicas de búsqueda, descubrimiento y administración de información para los usuarios. Debido a esto, es necesario conocer a los usuarios para poder brindarles información que les sea útil y relevante para sus diversas necesidades y contextos. La construcción de *perfiles* nos permite almacenar preferencias y/o intereses del usuario, tanto de corto como de largo plazo, con el objetivo de poder llevarle información relevante en el momento que lo requiera, así como adaptar los sistemas o interfaces a sus necesidades. Existen *repositorios de información personal* (Teevan *et al.*, 2005) que el usuario va forman-

[1] 1 Petabyte= 1,048,576 Gigabytes

[2] 1 Zettabyte= 1,048,576 Petabyte

do implícitamente en su interacción con la computadora e Internet; por ejemplo: las páginas web visitadas, páginas web guardadas (*los marcadores de Internet*, **Bookmarks**), las búsquedas realizadas, los correos electrónicos, notas en el calendario, documentos almacenados en la computadora, etc. Todas estas fuentes de información personal que el usuario generalmente guarda de su computador, los podemos utilizar para construir su *perfil*.

1.1 Motivación y justificación

Todos los sistemas y dispositivos con los cuales las personas interactúan pueden generar información de la actividad con los usuarios, la cual permite hacer observaciones y proponer herramientas o técnicas para una mejor interacción. El almacenamiento de esta información nos permite procesarla para explotarla y generar nuevas *interfaces, sistemas adaptativos, sistemas de recomendación, sistemas de búsqueda personalizada, sistemas de recuperación de información, sistemas de filtrado,* entre algunos ejemplos.

La construcción de perfiles nos permite guardar información del usuario para poder llevar a cabo este tipo de tareas. Podemos hacer que el usuario pueda tener una mejor experiencia no sólo con la navegación en Internet sino también con la interacción en sus demás dispositivos, independientemente de cuales sean. Por ejemplo, si el usuario adquiere un teléfono celular con sistema operativo **Android**[3] y tiene una cuenta de correo en **Gmail**[4], el usuario puede configurar su cuenta en el dispositivo móvil y con eso tener acceso a las preferencias de esta en el dispositivo— además de que el celular se convierte ahora en un medio de recolección de información que se agrega a la cuenta de correo configurada. Podemos, de igual manera, hacer que el usuario pueda acceder a sus preferencias entre diversos dispositivos, como sucede con

[3] www.android.com/about/

[4] www.google.com/gmail

la herramienta del navegador web **Firefox**[5] llamada **Firefox Sync**[6], la cual permite al usuario guardar contraseñas, marcadores (Bookmarks), historial de navegación y pestañas abiertas para ser utilizado en cualquier equipo de cómputo fijo o móvil que tenga instalado dicho navegador web y conexión a Internet.

Las preferencias del usuario, representadas en su perfil, nos permiten crear sistemas que se adapten a sus necesidades; como ejemplo podemos mencionar el sitio web de Internet **www.youtube.com** en donde los videos observados por el usuario una vez registrado en el sitio de Internet son almacenados para que la próxima vez que entre al sitio puedan aparecer videos de los mismos temas ya antes vistos. También podemos generar sistemas para filtrar información que sea relevante para el individuo como lo hace el sitio de Internet **Google** en su sección de *Google noticias*[7], en donde los usuarios previamente registrados con una cuenta de correo en **Gmail** pueden crear un tema y *Google noticias* se encargará de traer noticias relacionadas al tema.

La información que el usuario nos proporcione (*información explícita*) o vayamos conociendo de él con base a su interacción con el sistema o dispositivo (*información implícita*), la almacenamos en un perfil de usuario. El estudio de perfiles nos permite lograr llevar a los usuarios, con base a sus preferencias y/o intereses, información relevante y una mejor interacción con los sistemas y dispositivos.

1.2 Definición del problema

Con el crecimiento de datos tan grande que se ha estado generando en Internet, la vida diaria de los usuarios va integrando cada vez más herramientas, servicios y dispositivos que interactúan con esta red debido a la cantidad de información que se mueve en su interior. Las áreas del conocimiento que se dediquen al filtrado de estos datos, a la optimización de su presentación, a la búsqueda personalizada y al

[5] www.mozilla.org/es-ES/firefox/fx/

[6] www.mozilla.org/es-ES/mobile/sync/

[7] http://news.google.com.mx/

descubrimiento de nueva información, serán las encargadas de dar respuesta a los nuevos retos que se van a enfrentar.

La *personalización* es un mecanismo que nos permite llevar información que sea importante para los usuarios, así como la creación de *sistemas adaptativos* que se moldeen a las necesidades de los diferentes entornos que enfrentan los usuarios. Con la *generación de perfiles* podemos guardar la información que necesitamos para llevar acabo una personalización, sin embargo, nos encontramos con dos problemas que son: obtener la fuente de información del usuario, de donde vamos hacer la extracción de sus intereses y el segundo problema es el escenario de un repositorio escaso de información.

Nuestra investigación propone un mecanismo para generar perfiles de usuario a partir de repositorios de información personal así como un proceso de enriquecimiento que afronte el escenario de repositorios con escasa información.

1.3 Protocolo de investigación

Los siguientes puntos conformarán el protocolo de investigación: objetivo general, objetivos específicos, preguntas de investigación e hipótesis.

1.3.1 Objetivos

Objetivo general. *Generar un perfil de usuario a partir de un repositorio de información personal implícito.*

Objetivos específicos.

1. Tomar un repositorio con información del usuario y procesar su contenido.
2. Generar una representación del perfil del usuario a partir del repositorio.
3. Aplicar un mecanismo de enriquecimiento al perfil del usuario.

4. Generar una visualización del perfil del usuario.

1.3.2 Preguntas de investigación

- ¿Es posible construir un perfil de usuario a partir de un repositorio de información personal?
- ¿Cómo representar los intereses del usuario?
- ¿Es posible enriquecer el perfil del usuario a partir de su contenido?
- ¿Cómo visualizar el perfil del usuario?

1.3.3 Hipótesis

A continuación describiremos las hipótesis de esta investigación.

- Es posible construir un perfil de usuario a partir de un repositorio de información del usuario.
- Es posible enriquecer el perfil a partir de su contenido.
- Podemos hacer una representación visual del perfil.

El presente documento queda distribuido de la siguiente manera: en el Capítulo 2 expondremos los antecedentes que fueron explorados para llevar a cabo esta investigación, así como trabajos relacionados; en el Capítulo 3 hablaremos del modelo conceptual desarrollado. El Capítulo 4, presenta el caso de estudio de la investigación. En el Capítulo 5 mostraremos la experimentación desarrollada durante esta investigación y por último expondremos nuestras conclusiones y trabajos futuros en el Capítulo 6.

Capítulo 2

Antecedentes

Con el crecimiento que se ha estado teniendo con Internet (Gantz y Reinsel, 2010), los usuarios nos encontramos con panoramas cada vez más complejos para poder hacernos de la información que necesitamos y encontrar aquella que nos pueda ser útil. En estos tiempos si no sabemos la dirección de una página web, difícilmente la vamos a poder localizar a través de las búsquedas que llevemos a cabo a través de Internet. Son conocidos tres escenarios (Kosala y Blockeel, 2000) con los cuales se enfrentan los usuarios en Internet y son:

- Localización de información relevante.
- Creación de nuevo conocimiento a partir de la información que se encuentra en Internet.
- Personalización de la información.

En estos escenarios existen herramientas para ayudar a los usuarios y todas estas herramientas necesitan de su información para poder llevar a cabo sus tareas. A la información generada por el usuario, almacenada en un conjunto de datos o documentos podemos llamarla *Repositorio de información* y este repositorio de información puede ser procesado para generar perfiles de usuarios. Los perfiles nos ayudan a generar personalizaciones, recomendaciones y/o aplicar filtros a los sistemas con los cuales el usuario interactúe, para llevar información que este más cercana a sus necesidades (Gauch *et al.*, 2007).

2.1 Perfiles

El propósito de un perfil de usuario es mantener información de las preferencias (Gils y Schabell, 2003) o intereses (Pretschner, 1998) dentro de un repositorio que podemos utilizar para traer información relevante. Un perfil de usuario nos permite crear una personalización de la información o servicio para llevarle la información correcta que necesita en el tiempo que se le solicita (Speretta, 2000). Existen páginas web como *www.trap.it* y *www.stumbleupon.com* que le permiten al usuario guardar preferencias de temas, las cuales llevan a cabo una función de asistentes personales (P.R.Kaushik y Murthy, 1999) para hacer búsquedas automáticas de los temas guardados por el usuario. Estos sistemas requieren *información explícita* por parte del usuario; es decir, ocupan que este llene formularios con sus datos y preferencias. Los perfiles de este tipo de sistemas suelen ser *perfiles fijos*; o sea que si el usuario cambia de intereses con el tiempo, debe de ir a donde se encuentra su información y colocar en el sistema que ya prefiere otros temas, para que el sistema busque ahora los temas nuevos. Para recolectar *información explícita* es común que el usuario tenga que llevar a cabo algunas acciones, no sólo llenar formularios sino también instalando algún tipo de software como son los agentes de escritorio e.g. Google Desktop, agentes de navegación (Mladenic, 1996), etc. Los mecanismos que buscan *información implícita* de los usuarios son los que se convierten en un verdadero reto ya que el objetivo es hacerlo sin que el usuario intervenga, sin que configure o instale ningún tipo de software. A pesar de esta polarización en la búsqueda de las soluciones, también se han creado *sistemas híbridos* (Billsus y Pazzani, 1999) con la finalidad de contribuir al conocimiento.

Los perfiles que consideran el tiempo dentro de los intereses de los usuarios son *perfiles dinámicos* (Gauch *et al.*, 2007) y diferencian los intereses de corto plazo y de largo plazo de los usuarios. Algunos sólo consideran los intereses a largo plazo (Kuflik y Shoval, 2000) debido a que este tipo de intereses son los que pueden definir mejor el perfil.

El ciclo básico de la creación de perfiles de usuarios usualmente sigue un proceso de tres etapas (Gauch *et al.*, 2007):

1. Se hace acopio de la información del usuario.
2. Se toma la información recolectada para ser procesada y construir el perfil de usuario.
3. Se utiliza el perfil del usuario para brindar algún servicio personalizado.

Los mecanismos de recolección de *información explícita* son los formularios en papel o en web y las encuestas de lo que los usuarios prefieren; también existen mecanismos de recolección de *información implícita*, que son:

- **Servidores Proxy** (*Proxy Servers*). (Trajkova y Gauch, 2004)
- **Bitácoras Web** (*Web Logs*). (Mobasher, 2007)
- **Bitácoras de búsqueda** (*Search Logs*). (Sieg *et al.*, 2004)
- **Caché de navegador** (*Browser Cache*). (Pretschner, 1998)
- **Agentes de navegación** (*Browser Agents*). (Lieberman *et al.*, 1995)
- **Agentes de escritorio** (*Desktop Agents*). (Dumais *et al.*, 2003)

Diversas fuentes de información personal pueden ser utilizadas para la creación del perfil del usuario, Teevan *et al.* (2005) utilizaron un agente de escritorio para analizar páginas web visitadas, búsquedas realizadas, correos electrónicos, notas en el calendario y todos los documentos almacenados en la computadora. Experimentaron con diversas reglas para la extracción de la información así como la creación de perfiles de los cuales observaron su comportamiento, pero fue el perfil que integró toda la información del usuario el más preciso.

Uno de los primeros trabajos que utilizó el historial de navegación del usuario así como sus marcadores de Internet fue el realizado en Letizia (Lieberman *et*

al., 1995) en donde a partir de estas fuentes de información se extraía información implícita para hacer sugerencias de enlaces con temas de intéres para el usuario. Montebello *et al.* (1998) utilizaron marcadores de Internet para crear los perfiles de usuario y utilizaron la frecuencia de palabras para extrar los intereses, pero en la investigación se propuso cada marcador de Internet como un vector, de tal forma que el perfil del usuario era un grupo de vectores.

Los trabajos con marcadores de Internet no han sido ajenos a la influencia de las redes sociales y de cómo el Internet se vuelve más colaborativo. Debido a esto existen trabajos como el de Kanawati y Malek (2002); Jung (2005), entre otros, en donde a través de los marcadores de Internet de los usuarios se extraen intereses que son buscados en intereses de otros usuarios para poder recomendar material en común con el fin de ahorrar tiempo de búsqueda en Internet. Estos mecanismos pueden ser utilizados en grupos de investigación de algún tema en común o estudiantes que llevan a cabo tareas de búsqueda de información.

La representación de los perfiles se ha realizado con *palabras claves* con un peso que permita su ordenamiento (Moukas, 1996), *redes semánticas* (Gentilia *et al.*, 2003), *agrupaciones de conceptos* con un peso (Trajkova y Gauch, 2004) y/o *reglas de asociación* (Mobasher *et al.*, 2002).

2.2 Sistemas de recomendación

Los **sistemas de recomendación** (**Recommender Systems**, **RS**) nacen como un mecanismo que afronta las condiciones tan complejas en las que se encuentran los usuarios en Internet ante la cantidad tan grande de información. Así como las personas recomiendan a sus amistades o familiares películas, música, libros, eventos diversos, objetos diversos, etc. Los sistemas de recomendación buscan hacer lo mismo para llevar información a las personas que les pueda ser útil. Los sistemas de recomendación (Ricci *et al.*, 2011) los podemos definir como herramientas de software y técnicas que proveen sugerencias de objetos (eventos, artículos, página web, etc.)

que pueden ser de utilidad a los usuarios. La clasificación de estos sistemas según Balabanović y Shoham (1997) es:

- **Sistemas de recomendación basados en contenido**. (Pazzani *et al.*, 1996; Balabanović y Shoham, 1997)
- **Sistemas de recomendación colaborativa**. (Breese *et al.*, 1998; Konstan *et al.*, 1997)
- **Sistemas híbridos de recomendación**. (Pazzani, 1999)

En los **sistemas de recomendación basados en contenido** al usuario se le recomiendan objetos (eventos,artículos, página web, etc.) basados en votaciones que el usuario realizó en el pasado acerca de objetos similares. Estos sistemas de recomendación extraen información del usuario a través de la interacción que el usuario tenga con el sistema; por ejemplo: un usuario registrado en una página web de libros. La página web observará los libros visitados por el usuario y las búsquedas realizadas para poder hacer recomendaciones de libros sobre los temas relacionados, aquellas votaciones que el usuario haga sobre algunas recomendaciones serán agregadas a su perfil para, con el tiempo, hacer que las sugerencias sean más certeras. Estos sistemas de recomendación utilizan en su mayoría herramientas del área de *recuperación de información* para la extracción de información de los contenidos. Mecanismos como palabras por frecuencia (**TF**) y palabras por peso (**TFIDF**) son de los muy utilizados; de igual manera, para extraer la relación entre los objetos el mecanismo más utilizado es la similitud cosenoidal (**Cosine Similarity**). También existen investigaciones en otras áreas como la inteligencia artificial (Pazzani y Billsus, 1997), aprendizaje de máquina (Billsus *et al.*, 2000) y métodos probabilísticos (Pazzani *et al.*, 1996).

Los **sistemas de recomendación colaborativa** son sistemas que tratan de predecir la relevancia de un objeto (evento,artículo, página web, etc.) para un usuario, en los objetos votados previamente por otros usuarios relacionados con él. Estos

sistemas de recomendación pueden ser agrupados según Breese *et al.* (1998) en dos grupos:

- Los **basados en memoria**. (Delgado y Ishii, 1999)
- Los **basados en el modelo**. (Marlin, 2003)

Los *sistemas de recomendación colaborativa basados en memoria* son aquellos que requieren todas la votaciones de los objetos, todos los objetos y todos los usuarios, cargados en memoria para hacer las sugerencias. Por otra parte los *sistemas de recomendación colaborativa basados en el modelo*, a través de algoritmos crean un modelo que utiliza un resumen de las votaciones para detectar patrones con las cuales hacen sugerencias.

Ambos modelos de sistemas de recomendación, tanto los *basados en contenido* como *los colaborativos*, sufren de dos principales complicaciones: la más importante es el de la falta de información para hacer las recomendaciones (conocido como nuevo usuario y nuevo objeto) y el segundo problema son los usuarios con intereses muy fuera del promedio de la población. Para aportar soluciones a estos dos tipos de escenarios, fueron generados los mecanismos híbridos.

Los **sistemas de recomendación híbridos** (Pazzani, 1999; Balabanović y Shoham, 1997; Schein *et al.*, 2002) buscan combinar diferentes técnicas o métodos de sistemas de recomendación conocidos, para buscar compensar carencias de otros sistemas de recomendación.

Se han identificado siete tipos de sistemas de hibridación (Burke, 2002), los cuales son:

- Pesos (*Weighted*) (Claypool *et al.*, 1999). El resultado de diferentes componentes es combinado.
- Intercambio (*Switching*) (Billsus y Pazzani, 2000). El sistema elige entre varios componentes de recomendación y dependiendo del criterio elige el componente.

- Mezcla (*Mixed*) (Smyth y Cotter, 2000). Recomendaciones de diversos sistemas de recomendación son utilizados juntos.

- Combinación de características (*Feature Combination*) (Basu *et al.*, 1998). Se utilizan características extraídas de diversas fuentes de información y son combinadas para asignarse a un solo algoritmo de recomendación.

- Aumento de características (*Feature Augmentation*) (Mobasher *et al.*, 2004). Una técnica de recomendación es utilizada para producir un atributo o un grupo de atributos que se convierten en la entrada para otra técnica.

- Cascada (*Cascade*) (Burke, 2002). Los que recomiendan más son tomados más en cuenta que los que recomiendan menos.

- Meta-nivel (*Meta-level*) (Balabanović y Shoham, 1997). Se crea un modelo a partir de un sistema de recomendación para que este sea la entrada a otro sistema.

A pesar de que en la clasificación de Balabanović y Shoham (1997) no mencionan los *sistemas de recomendación basados en reglas* estos mecanismos existen; por ejemplo: en la página web *www.stubleupon.com* se crea un perfil en donde se establecen los temas que el usuario tiene por preferencia, para que el mecanismo de búsqueda de la página traiga contenido acerca de ellos. Una vez que se tengan estos contenidos, ya sean videos, páginas web, imágenes, etc. *www.stubleupon.com* coloca unos botones para que el usuario marque si le gustó el contenido o no. Si el usuario da su voto esta información se coloca en el perfil del usuario, de tal forma que las próximas sugerencias serán hechas más cercanas a las marcadas positivamente y más lejanas de las marcadas negativamente. La página web *www.trap.it* también tiene un sistema similar pero sólo funciona con páginas web que contienen los temas que el usuario marcó en su perfil, en ambas páginas web si el usuario no utiliza el mecanismo de votación para señalar lo que le gustó o no, estas serán sólo agentes de búsqueda. Según Mobasher (2007) los *sistemas de recomendación basados en reglas*

tienen la característica de ser mecanismos que parten de un repositorio de información explícita como pueden ser la *información demográfica* del usuario (edad, genero, tamaño, peso, etc.), *información psicográfica* (clase social, estilo de vida, personalidad, gustos, etc.), así como algunas otras características del usuario; para hacer las recomendaciones.

2.3 Minería de texto

Según Feldman y Sanger (2006) la *minería de texto* la podemos definir como el proceso de descubrir información en una basta colección de texto y automáticamente identificar patrones de interés y relaciones en el texto.

Para poder llevar a cabo la tarea de la *minería de texto*, es necesario contar con una colección de documentos para llevar a cabo una preparación antes de su aplicación. En esta área es común ver a los documentos de texto como una representación de agrupamiento de palabras o *sacos de palabras* (**BoW**, *Bag of Words*).

Los procesos básicos para la preparación de los documentos de texto son (Hotho *et al.*, 2005):

- **Tokenización** (*Tokenization*).
- **Filtrado de palabras** (*Word Filtering*).
- **Lematización** (*Lemmatization*).
- **Derivar** (*Stemming*).

El proceso de **Tokenización** (*Tokenization*) consiste en eliminar del documento todos los símbolos de puntuación y sustituir tabulaciones y caracteres que no son texto por espacios en blanco, una vez realizado este procedimiento la colección de palabras que queda se le conoce como *diccionario*.

El siguiente proceso es el **Filtrado de palabras** (*Word Filtering*) del *diccionario*. Un listado básico de palabras a eliminar es el conocido como *palabras vacías* (**StopWords**) que consta de palabras que son artículos, preposiciones, conjunciones, etc. En este proceso pueden también agregarse *listas negras* (**Black lists**) que son grupo de palabras que se consideren a eliminar según experimentación.

Lematización (*Lemmatization*) es el proceso en donde se lleva a cabo la conversión de los verbos al tiempo infinitivo y busca pasar los sustantivos a su forma singular (*e.g. andabamos - andar*).

Derivar (*Stemming*) es el proceso de transformar las palabras a su forma básica sin plural o sin afijos (prefijos, sufijos, infijos, etc.). Los trabajos de Porter *et al.* (1980) son los más destacados para el *Stemming*.

En la *minería de texto* los documentos no siempre son tomados como agrupamientos de palabras sin tener en cuenta la semántica, en ocasiones preprocesamiento lingüístico (Manning y Schütze, 1999) se lleva a cabo para mejorar la información de los términos en los documentos. Existen cuatro métodos:

- **Part-of-speech tagging** (**POS tagging**). Agrega la relación de las palabras con sus palabras adyacentes y las etiqueta.
- **Text chunk**. Divide un texto en grupos de palabras relacionadas sintácticamente.
- **Word Sense Disambiguation**. (**WSD**) Trata de resolver la ambigüedad entre las palabras.
- **Parsing**. Produce un árbol de análisis de la oración para poder observar la relación de cada palabra con el resto.

2.3.1 Minería de texto y recuperación de información

La *minería de texto* y la *recuperación de información* (**Information Retrieval**, **IR**) tienen una relación cercana como áreas de investigación. *La recuperación de información* es encontrar el material (generalmente documentos) de una naturaleza no estructurada (generalmente texto) que satisface una necesidad de información desde dentro de grandes colecciones (normalmente almacenada en las computadoras) (Manning *et al.*, 2008). Los mecanismos utilizados para hacer el procesamiento de las colecciones de datos se llevan a cabo con los procesos o técnicas de la *minería de texto* mostrados anteriormente, si vamos a preparar la colección de documentos para hacer una evaluación de los pesos de sus palabras con respecto a la colección o con respecto al documento, si necesitamos conocer la relación que guardan con respecto a otras colecciones de documentos o con respecto a otro mismo documento de la colección, etc. Esto lo podemos llevar a cabo con conocimiento del área de *recuperación de información* como el **TFIDF**, **Similitud Cosenoidal**, **Modelo Espacio Vectorial** entre otros mecanismos.

Se ha propuesto una categorización de los modelos téoricos básicos de la recuperación de información (Dominich, 2000) los cuales son:

- **Modelos clásicos**. Los tres modelos clásicos son: Boleano, Modelo Espacio Vectorial y el Probabilístico.
- **Modelos alternativos**. Estos modelos están basados principalmente en técnicas de teoría de conjuntos difusos.
- **Modelos lógicos**. Son aquellos basados en lógica formal.
- **Modelos basados en interactividad**. Son aquellos que toman los elementos de la colección de datos como agrupamientos fuertemente interconectados.
- **Modelos basados en inteligencia artificial**. Son aquellos que utilizan principalmente métodos y técnicas del área de inteligencia artificial como son: *redes neuronales*, *algoritmos genéticos*, *procesamiento del lenguaje natural*, etc.

La *recuperación de la información* es un área de investigación muy amplia con técnicas multidiciplinarias y en esta sección mencionaremos brevemente algunos métodos utilizados en esta investigación, como son: *modelo espacio vectorial, TFIDF y similitud cosenoidal.*

2.3.2 Modelo Espacio Vectorial

El modelo espacio vectorial (Manning *et al.*, 2008) es un modelo que expresa los documentos como vectores y da a las palabras de los documentos, una representación de peso para su ordenamiento. Consideremos los siguientes documentos para hacer un ejemplo del modelo.

- Documento 1: las computadoras están incrementando su potencial.
- Documento 2: computadoras de escritorio menos compradas durante navidad.
- Documento 3: los escritorios incrementan su costo.

A estos documentos aplicamos las técnicas de *minería de texto* para el procesamiento, como se muestra a continuación.

	computadora	incrementar	potencial	escritorio	comprar	navidad	costo
Doc 1	1	1	1	0	0	0	0
Doc 2	1	0	0	1	1	1	0
Doc 3	0	1	0	0	1	0	1

Tabla 2.1: Representación matricial del modelo espacio vectorial

Como se puede observar en la Tabla 2.1, han sido eliminadas las *palabras vacías* (las, están, su, de, menos, durante, los), la palabra *computadoras* ha sido cambiada su forma singular y las palabras *incrementando* e *incrementan* han sido pasadas a su forma raíz. Ahora cada documento queda representado en forma de vector:

- Documento 1 = (1, 1, 1, 0, 0, 0, 0)

- Documento 2 = (1, 0, 0, 1, 1, 1, 0)
- Documento 3 = (0, 1, 0, 0, 1, 0, 1)

Si quisiéramos llevar a cabo una consulta para recuperar información es necesario convertir esa consulta utilizando la matriz de la Tabla 2.1; por ejemplo: si hiciéramos una consulta que buscara la frase *Nuevas computadoras de escritorio* en la colección de documentos. Debemos representar la consulta como un vector $\vec{V}$ = (1, 0, 0, 1, 0, 0, 0) para poder buscar la similitud entre la colección de documentos y traer los más relevantes organizados de mayor a menor. *El modelo espacio vectorial* contiene documentos uniformes en su contenido pero si uno o varios documentos tuvieran la palabra computadora muchas veces, la posición de los resultados se vería afectada. De tal forma que es necesario utilizar un mecanismo de normalización.

La utilización del $TF * IDF$ (Manning *et al.*, 2008) (**Term Frequency - Inverse Document Frequency**) nos ayuda a afrontar este escenario, debido a que la *frecuencia de las palabras* es considerada en el documento ($TF_{i,j}$) y la *frecuencia inversa del documento* (IDF_i) actúa como discriminador del término popular en un documento ante la colección completa.

$$(TF * IDF)_{i,j} = W_{i,j} = tf_{i,j} \times log\frac{N}{idf_i} \tag{2.1}$$

$(tf * idf)_{i,j} = W_{i,j}$ = representa el peso de i en j

$tf_{i,j}$ = número de ocurrencias de i en j

idf_i = número de j que contienen i

N = número total de j

2.3.3 Similitud Cosenoidal

La escencia de la *recuperación de información* es llevar a cabo la tarea de comparar una búsqueda con una colección de documentos para regresar un ordenamiento

cercano a la búsqueda realizada. La manera en que se llevó a cabo la comparación de la relación que los repositorios de usuario tenían con respecto al resto de sus compañeros, fue a través de la *similitud cosenoidal* (Manning *et al.*, 2008). Para un par de documentos, se mide la similitud a través del coseno del ángulo, de tal forma que:

$$Similarity = Cos(\theta) = \frac{\vec{X} \cdot \vec{Y}}{\| \vec{X} \| \| \vec{Y} \|} \tag{2.2}$$

Los mecanismo utilizados durante el desarrollo de la investigación como *TFIDF* y *similitud cosenoidal* fueron considerados, basados en investigaciones anteriores, que serán mencionadas en la siguiente sección de trabajos relacionados.

2.4 Trabajos relacionados

Diversas fuentes de datos (Teevan *et al.*, 2005) del usuario han sido utilizadas para la creación de perfiles que permitan brindar mecanismos de filtrado o personalización a los usuarios, conforme el crecimiento que se está dando en Internet. En esta investigación se tomó la decisión de utilizar los marcadores de Internet, como en Montebello *et al.* (1998) que los utilizaron para crear los perfiles de usuario. En la investigación usaron la frecuencia de palabras para extraer los intereses y se propuso cada marcador de Internet como un vector, de tal forma que el perfil del usuario era un grupo de vectores. Esta forma de utilizar la representación de los marcadores de Internet es debido a que en un conjunto de documentos que es representado como un solo vector, la probabilidad de similaridad es menor que si se busca un vector dentro del grupo de vectores que contenga la similitud. Los trabajos con los marcadores de Internet no se hicieron ajenos a la evolución de Internet y empezaron a ser influnciados por la tendencia social y de colaboración que se ha estado dando. De tal forma que investigaciones como Kanawati y Malek (2002); Jung (2005) utilizan los marcadores de los usuarios para sugerir contenido a los integrantes de estos grupos de colaboración. La página web *delicious.com* es una red social que basa la creación

de relaciones en un servicio de gestión de marcadores de Internet, los cuales representan las preferencias de los usuarios y son estas preferencias las que hacen que los usuarios se relacionen entre ellos. Esta página web no solo almacena los enlaces de los usuarios, también permite a los usuarios generar agrupamientos de categorías para clasificar sus contenidos que son enlaces a grupos de usuarios con mismas preferencias. Este mecanismo de etiquetas generadas por los usuarios a partir de sus preferencias es llamado *Folksonomies* (Bogárdi-Mészöly *et al.*, 2013) y ha sido bien estudiado (Sinclair y Cardew-Hall, 2008). Nuestra propuesta para la visualización de los perfiles a través de nubes de palabras es generada a partir de los marcadores de Internet de los usuarios, pero extraídos sin intervención de los usuarios.

WordNet ha sido y sigue siendo aplicado en muchas investigaciones por su gran aporte al conocimiento de la lengua con su estructura. En perfiles ha sido utilizado para aportar a la información de los usuarios como Semeraro *et al.* (2005) en donde emplearon **WordNet** para enriquecer el perfil de los usuarios seleccionando *sinónimos* e *hiperónimos*. Los *sinónimos* fueron utilizados para eliminar la *polisemia* de los perfiles de los usuarios y los *hiperónimos* para buscar darles contexto a las palabras y más precisión. En esta investigación Semeraro *et al.* (2005) usan un mecanismo automático con técnicas de aprendizaje de máquina y desarrollan su investigación con el principio empleado en Montebello *et al.* (1998) en donde también crean el repositorio del usuario como un grupo de vectores al cual llaman *Bag of Sets* (saco de conjuntos). También en Bouras y Tsogkas (2012) fueron aplicadas las relaciones de **WordNet** de *hiperónimos* para enriquecer el perfil de usuario y posteriormente utilizar la información para un sistema de recomendación, la técnica para representar el perfil fue de un cluster de conceptos. **WordNet** también ha sido utilizada no sólo para darle contexto al contenido de los perfiles sino también al de documentos, que pueden ayudar a una mejor categorización como Rodriguez *et al.* (2000) y Scott *et al.* (1998). Ambos trabajos son un ejemplo de cómo podemos aplicar **WordNet** como un trabajo previo a la creación en los documentos de los usuarios para poder crear perfiles con contenido previamente categorizado a través de **WordNet**. Nuestra propuesta extrae las palabras con la relación de *hiperónimos*

en los *synset* pero no usamos método alguno para la desambigüedad debido al tamaño de los repositorios de los usuarios, también nuestra propuesta busca afrontar el fenómeno de repositorios pequeños.

Pero **WordNet** no es la única estructura que se utiliza en estas investigaciones, también ha sido utilizada **Wikipedia**[1] como en Ramanathan *et al.* (2008) que fue utilizada para crear una jerarquía en el perfil del usuario que fueron extraídas de los conceptos de la estructura de **Wikipedia.** Bajo este mecanismo también es posible hacer un aporte ante los problemas que causa la ambigüedad Jabeen *et al.* (2012) y la falta de contexto en los conceptos en los pefiles de usuario, pero el principal problema es el llevar a cabo la búsqueda de los conceptos en el índice de **Wikipedia** que en la investigación era de 1.4 gigabytes.

2.5 Resumen

Durante este capítulo hicimos un repaso de los antecedentes de las áreas del conocimiento que abarca nuestra investigación. Vimos algunas herramientas de recuperación de información (*Information Retrieval*) así como de minería de texto. También repasamos el área de los sistemas de recomendación debido a que nuestro trabajo es un paso previo a este tipo de mecanismos.

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

Capítulo 3

El Modelo

En esta sección explicaremos cómo se llevó a cabo el desarrollo de la investigación. En el primer punto se presenta el *marco formal de trabajo* en donde abarcamos de forma general los principales conceptos; como segundo punto, hablaremos de la *generación del perfil del usuario* en donde abordamos de forma más detallada su proceso de creación y como tercer punto exponemos cómo llevamos a cabo la *visualización del perfil* así como sus posibles aplicaciones y por último, hablaremos del *enriquecimiento del perfil* del usuario en donde abordaremos un mecanismo que nos ayude ante repositorios pequeños.

3.1 Marco formal de trabajo

La interacción que el usuario lleva a cabo con sus dispositivos electrónicos (*computador de escritorio, computador pórtatil, tableta, teléfono inteligente, TV inteligente, consola de videojuegos, etc.*), va dejando información que puede ser extraída para ser analizada. Datos como *páginas web visitadas, páginas web guardadas, cookies, documentos abiertos, documentos creados, correos electrónicos, sesiones abiertas, sistemas de mensajería, programas abiertos, etc.* forman parte del conjunto de datos que pueden ser reunidos para formar un repositorio de datos que nos permita conocer al usuario para poder brindarle soluciones a su medida; de tal forma que un usuario es visto en términos de un *repositorio de documentos* o como un *agrupamiento de datos*. Un documento, por su parte, es considerado simplemente como un

saco de palabras (**BoW**, *Bag of Words*) donde el orden de las mismas no es tomado en cuenta ni su gramática. De todas las palabras que existen en un documento dado, sabemos que algunas de ellas son las más significativas o sobresalientes y son estas palabras las que representan los temas que podemos tomar como intereses del usuario. Es entonces que el perfil de un usuario sería un meta-documento que incluye las palabras más sobresalientes de cada documento del repositorio personal.

Podemos representar de manera formal el repositorio del usuario u como el conjunto $R_u = \{D_1 \dots D_n\}$ donde n representa la cantidad de documentos del repositorio y $D_i = \{t_1 \dots t_r\}$ es un documento con r palabras.Un documento filtrado $\delta_i \subseteq D_i$ contiene sólo aquellas palabras sobresalientes, de acuerdo a una función $f(i, t_j) \to \mathbb{R}$. Esta permite evaluar cada palabra t_j para el documento i y, así, escoger las k palabras más significativas (note que $|\delta_i| = k$). Un perfil de usuario, por tanto, resulta en la unión de los documentos filtrados del repositorio R_u, de tal suerte que:

$$P_u = \bigcup_{1}^{n} \delta_i$$

En el modelo, se contemplan, además dos operaciones sobre el perfil: *visualización* y *enriquecimiento*.

El *enriquecimiento del perfil* consiste en agregar palabras desde una fuente externa de conocimiento como sería un diccionario, ontología o recurso léxico—entre otros. Esta agregación de palabras sería horizontal, es decir, con respecto al contexto (*e.g. language, linguistic communication, art, artistic creation*) o vertical, es decir, con respecto a la estructura del lenguaje (*e.g. sinónimos: language, speech, voice communication, oral communication*).

Hablando formalmente, si el conjunto Ψ representa el vocabulario (serie de palabras únicas) de la fuente externa, el perfil enriquecido denotado por E_u consistiría en la unión de un subconjunto de la fuente $\psi \subset \Psi$ con el perfil actual:

$$E_u = P_u \cup \psi$$

De hecho, las palabras que enriquecen el perfil están relacionadas con el perfil mismo. En ese sentido, ψ es la proyección de una relación $R \subset P_u \times \Psi$:

$$\psi = \pi_{\Psi}(R)$$

La *visualización* consiste en asignar atributos a las palabras del perfil, tales como *tamaño, color* y *posición* (Tabla 3.1). Estos atributos están en función de la relevancia que las palabras tienen en el repositorio del usuario, esa relevancia es extraída bajo diversas reglas que el investigador selecciona dependiendo de cuál sea su objetivo. Para ello, podríamos considerar que se tiene una matriz $\vec{V_u} = \vec{T_u}\vec{A}$, donde la casilla v_{ij} representa el valor del atributo a_j para la palabra t_i.

	R	**G**	**B**	x	y	*Tamaño de letra*
Computer	255	110	35	40	20	30
Science	60	200	48	50	20	24

Tabla 3.1: Ejemplo de una visualización que incluye atributos para color (RGB), posición en 2D (x,y) y tamaño de letra.

3.1.1 Procesamiento del repositorio

Como lo mencionamos anteriormente, la interacción del usuario con diversos dispositivos electrónicos deja datos que nos pueden dar información que podemos utilizar. Los datos son dejados por el usuario en diversas formas y dependiendo de los objetivos que tengamos, es la manera en la que vamos a buscar la información que requerimos. Para reunir los diversos datos del usuario y formar su repositorio personal de información, es necesario seguir algunos pasos:

Recopilación del repositorio.- Consiste en determinar la fuente o las fuentes de

extracción para juntar las piezas que conformarán el repositorio de información personal.

Limpieza.- Pre-procesamiento de las piezas de información, de tal forma que queden como *sacos de palabras* (repositorio R_u). Pasos de la minería de textos (Feldman y Sanger, 2006) que son aplicados durante este proceso pueden ser *tokenización*, *filtrado de palabras*, *lematización* y *derivación*.

Generación del perfil.- Construir el perfil P_u a partir del repositorio de documentos.

Operaciones sobre el perfil.- Estas operaciones, como ya se mencionó, incluyen la visualización (generación de $\vec{V}$) y/o enriquecimiento del perfil (generación de E_u).

La recopilación del repositorio del usuario así como la limpieza del mismo, son temas que serán abordados con más detalle en el Capítulo 4.

3.2 Generación del perfil

Para iniciar con la construcción del perfil de usuario, el repositorio se encuentra como un grupo de documentos de texto, con signos de puntuación eliminados así como palabras vacías (**StopWords**) y listas negras (**Blacklists**). Hasta este punto los documentos están limpios para ser procesados en donde hay tópicos, palabras claves y/o palabras de peso por descubrir y que nos dan información acerca del usuario. Pero mientras las palabras más utilizadas nos pueden dar un resultado correcto al interior de un documento, esa misma frecuencia nos puede dar un resultado engañoso con respecto al repositorio completo del usuario.

Este modelo sugiere un análisis que nos permita tomar en cuenta la frecuencia de las palabras en un documento pero que nos permita castigar esa frecuencia con respecto al repositorio completo; es decir, tenemos que normalizar el peso de una

palabra dentro de su documento, de manera local, y el peso de la palabra dentro del repositorio completo, de manera global.

Elegimos utilizar el **TF*IDF** (Ecuación 3.1) (*Term Frequency Inverse Document Frequency*) para este propósito, siendo i la palabra y j el documento, en donde:

$$\mathbf{TF}_{i,j} = \frac{\text{número de veces que } i \text{ aparece en } j}{\text{el total de } i \text{ en } j}$$

$$\mathbf{IDF}_i = log\left(\frac{\text{número total de } j \text{ en el repositorio}}{\text{el número de } j \text{ en donde aparece } i}\right)$$

$$(TF * IDF)_{i,j} = W_{i,j} = tf_{i,j} \times log\frac{N}{idf_i} \tag{3.1}$$

$(tf * idf)_{i,j} = W_{i,j} =$ representa el peso de i en j

$R_u = N =$ número total de j

La multiplicación de estos valores **TF*IDF** nos da un valor que significa lo importante que es una palabra (i) para un documento (j) en el repositorio (R_u) del usuario analizado.

Generamos los valores de **TF** de los documentos a procesar, dando como resultado documentos de texto en un inicio como los mostados en la Tabla 3.2, en donde colocamos los valores de la frecuencia de las palabras al frente de cada palabra y ordenadas de mayor a menor frecuencia.

Una vez teniendo los valores de la frecuencia de la palabra dividimos entre el total de palabras de cada documento para obtener el valor de **TF**. La Tabla 3.3 muestra una representación parcial de un documento con el **TF** calculado.

Realizamos la operación de dividir el *número total de documentos del repositorio* entre *el número de documentos en donde aparece cada palabra*, a este resultado, sacamos el valor de logaritmo y obtenemos el valor **IDF** de cada documento. Te-

7 cfe
6 services
3 business
2 tinder
2 thousand
2 telecom
2 suppliers

Tabla 3.2: Muestra parcial de un documento previo al cálculo de TF.

0.038 cfe
0.033 services
0.016 business
0.011 tinder
0.011 thousand
0.011 telecom
0.011 suppliers

Tabla 3.3: Resultados TF de documento procesado.

niendo los valores **TF** e **IDF**, hacemos la multiplicación **TF*IDF** teniendo como resultado una lista ordenada de mayor a menor con los resultados por palabra. De la lista con los valores tomamos sólo las primeras 20 palabras por documento que representan los pesos más importantes.

En la Tabla 3.4 es una muestra breve de un documento de texto procesado con los valores del **TF*IDF**.

cfe 0.143
telecom .041
hydroelectric 0.041
commitments 0.041
services 0.033
responsibility 0.033
museum 0.033

Tabla 3.4: Documento procesado con TFIDF.

En este punto de la investigación tenemos por usuario, un grupo de documentos de texto procesado con los valores **TFIDF** y por documento su *top 20* de palabras que reflejan los temas relevantes. Para la generación de la visualización del perfil es necesario contar con un repositorio de usuario con una sola fuente de información. De tal forma que es necesario colocar el grupo de documentos del usuario a un solo documento, en donde vamos a extraer los temas que mostraremos en la visualización.

3.3 Visualización del perfil

Una visualización del perfil puede servir para que el usuario pueda tener accesos de su material ordenado por los temas extraídos, puede hacer una navegación sobre material nuevo de Internet a través de un filtrado de su visualización, entre algunas otras opciones.

El mecanismo de visualización del perfil del usuario, lo representamos como

una *nube de palabras* como las que se han popularizado en Internet a partir de la **Web 2.0**[1]. Las nubes de palabras (**Tag Cloud** en inglés que se refiere a *nube de etiquetas*) son comunes en algunas páginas de Internet en donde se muestran los temas más populares en la página web. En la Figura 3.1 se muestra un ejemplo encontrado en Internet (*http://wordpress.org/support/*) y podemos observar cómo los tamaños de las palabras son distintivos.

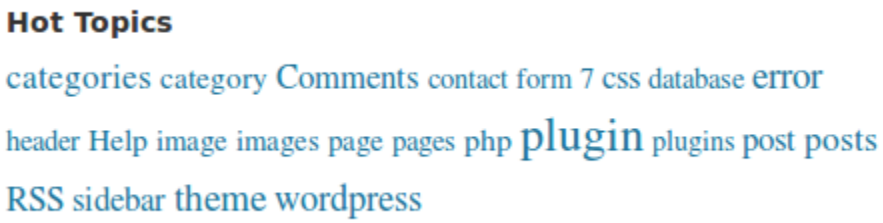


Figura 3.1: Ejemplo de Nube de etiquetas.

La razón que tomamos para utilizar este tipo de visualización es el hecho de que las palabras por sí solas pueden dar información, pero la manera en que son representadas le agregan valor adicional. Aún y que todas las palabras que aparecen en la representación visual del perfil son intereses del usuario, podemos observar cómo algunos intereses predominan más que otros.

Figura 3.2: Nube de palabras procesadas mediante TFIDF.

[1]http://tinyurl.com/bxaqdsj

La Figura 3.2 es un ejemplo de representación visual de uno de los usuarios procesados, podemos darnos cuenta cómo el usuario tiene intereses bien definidos en el área de computación o informática especialmente en el área de lenguajes como *HTML, Java* y *Python*. La palabra *example* destaca más que palabras como *tutorial, education, training, learning* y *howto*, pero con palabras como estas podemos darnos cuenta que en este repositorio existen páginas con un interés de aprendizaje muy probablemente sobre los lenguajes mencionados o en alguna del resto de las palabras.

Tenemos palabras como *linux, unix* y *gnu* que nos pueden dar un claro ejemplo de interés en sistemas operativos de código abierto. Palabras como *regular, expression, regex* y *regexp* nos muestran un interés en expresiones regulares que son utilizadas para la localización de patrones en palabras; si a eso agregamos palabras como *line, string, text, words, pattern, match* podemos ver como estas palabras también están relacionadas a expresiones regulares. También palabras como *mobile, android, apps* y *applications* nos muestran un interés, aunque más pequeño, acerca de aplicaciones para dispositivos móviles con la plataforma Android.

Podemos ver cómo la visualización del perfil nos deja información que nos puede ayudar a sugerir temas al usuario de forma más certera o podemos encontrar usuarios con los mismos intereses o similares, y tal como lo hacen las redes sociales hoy en día, podemos recomendarlos con otros usuarios para generar alguna amistad. En el Capítulo 6 hablaremos de trabajos que podemos llevar a cabo para continuar con las investigaciones en esta área.

Como mencionamos anteriormente durante este trabajo, existen varios problemas con los repositorios de los usuarios al momento que deseamos analizarlos para la búsqueda de intereses y un problema es el tamaño. En cuanto tengamos más información para procesar, los resultados de los perfiles salen con palabras como el de la Figura 3.2 en donde existen palabras que por su peso sobresalen entre las demás. En la Figura 3.3 mostramos un ejemplo de una nube de palabras de un usuario cuyo repositorio es muy pequeño.

Figura 3.3: Nube de palabras procesadas mediante TFIDF con repositorio pequeño.

Como podemos observar en la Figura 3.3, los tamaños de las palabras son uniformes. Este comportamiento se detectó en repositorios pequeños, en donde los usuarios no tenían muchas páginas para ser procesadas. Aun así podemos observar que existen palabras como *javascript, python, ruby, programming* y *programmers* que nos señalan un interés del usuario por la programación a través de estos lenguajes. También palabras como *learn, student, academic* y *codecademy* nos pueden indicar un interés por el aprendizaje de estos lenguajes o de sistemas operativos como *ubuntu* y *lubuntu*, que también se encuentran en la visualización.

Estas representaciones visuales como el de la Figura 3.3 han sido procesadas con el *TF*IDF* al igual que el resto de los repositorios, lo cual nos indica la normalización de las palabras pero aún y cuando sus palabras se ven uniformes en tamaño, estas nos dan información que podemos utilizar. Sin embargo, no podemos decir cuáles intereses son más fuertes que los demás debido a su uniformidad pero con repositorios mayores como el de la Figura 3.2 podemos decir cuáles intereses tienen mayor importancia que el resto, debido a que su peso en la visualización es evidente.

3.4 Enriquecimiento del perfil

Teniendo el dilema de los repositorios de usuario con pequeñas cantidades de información para ser analizada, es importante encontrar algún mecanismo que nos pueda ayudar a enriquecer el perfil del usuario. Llamamos *enriquecimiento del perfil* a la acción de agregar contenido de una entidad ordenada a otra, para incrementar su valor en información. Podemos considerar los siguientes pasos como pasos básicos a seguir en el enriquecimiento de perfil:

- Se revisan los objetivos que se quieren cumplir con el enriquecimiento para poder crear las reglas de extracción.
- Se revisan las entidades que puedan ayudarnos a cumplir los objetivos.
- Una vez hecha la elección de la entidad, se extraen los contenidos.
- Con los contenido extraídos, estos son agregados al repositorio que deseamos enriquecer.

Existen diversas estructuras ordenadas que pueden ser utilizadas en una investigación para su análisis. Podemos mencionar algunas como; *wiki.freebase.com* la cual es un grafo de conocimiento, *Wikipedia.org* es una entidad de conocimiento enciclopédico, *wordnet.princeton.edu* es una entidad semántica, entre otras más. Todas estas estructuras pueden ser utilizadas para enriquecer alguna información que tengamos; por ejemplo: si tenemos el nombre de un personaje histórico utilizando *wiki.freebase.com* podemos obtener información adicional como fecha de nacimiento, lugar de origen, personajes que se relacionen con él etc. Con *Wikipedia.org* podemos ampliar la información de tópicos y a través de los enlaces que nos recomiendan, podemos tomar tópicos que se relacionan ampliando la información. Con *wordnet.princeton.edu* podemos hacer una extracción de palabras que estén semánticamente relacionadas con alguna palabra de nuestro repositorio a analizar.

WordNet[2] es una base de datos léxica del idioma inglés y es administrada por la *Universidad de Princeton*. La estructura de **WordNet** está conformada por agrupaciones de palabras sinónimas, las cuales son llamados *synsets* (*synonym sets* o conjuntos de sinónimos) y tiene un total de 117000 *synsets* relacionados entre sí. Los tipos de palabras que guardan los *synsets* son verbos, sustantivos, adjetivos y adverbios. Estos tipos de palabras tienen relaciones semánticas con las cuales los *synsets* se conectan y son: *Hiperónimos, Hipónimos, Merónimos* y *Holónimos*. **WordNet** representa una gran ventaja para todos aquellos investigadores que deseen hacer experimentación utilizando base de datos léxica del idioma inglés, por su dimensión, su reconocimiento y que esta disponible para su utilización desde Internet (*http://wordnet.princeton.edu/wordnet/download/*). Así que nos dimos a la tarea de experimentar con su estructura y nuestros perfiles.

Hiperónimo	Flor es *hiperónimo* de Rosa, Clavel y Jazmín
Hipónimo	Rosa es *hipónimo* de Flor
Holónimo	Mano es *holónimo* de dedo y brazo
Merónimo	Dedo es *merónimo* de mano

Tabla 3.5: Algunas relaciones con las que WordNet asocia synsets.

Como podemos observar en la Tabla 3.5 el ejemplo de *hiperónimo* y el de *holónimo* pudieran parecer igual, pero no es así. La diferencia radica en el hecho de que en *hiperónimo* la rosa, el clavel o el jazmín, van hacer siempre una flor. Pero en el caso del *holónimo*, el dedo y el brazo, nunca van hacer una mano, pero forman parte de ella.

Si podemos localizar en los perfiles de los usuarios palabras que estén en los *synsets* de **WordNet**, entonces podemos tomar las relaciones con las que cuentan estas palabras para poder agregarlas a los perfiles de los usuarios y poder enriquecer el perfil. Así que, a través de un mecanismo de búsqueda de palabras, localizamos en **WordNet** las palabras que se encuentran en los perfiles de los usuarios y empezamos a generar un archivo de texto ahora con las palabras del perfil más las palabras

[2]http://wordnet.princeton.edu/

extraídas de **WordNet**. La relación que tomamos en cuenta para la extracción de las palabras de los *synsets* fue la de *hiperónimos*.

El programa toma una palabra del perfil del usuario y la busca en **WordNet**, al localizarla guarda sólo las palabras que se encuentren en el *synset* con la relación que se tomó y vuelve a repetir el procedimiento para cada palabra del perfil. Esto nos da algunas palabras repetidas, pero es ese fenómeno el que nos va a dar los pesos en la representación visual.

Figura 3.4: Nube de palabras procesadas mediante TFIDF y enriquecida con WordNet.

La Figura 3.4 muestra el resultado del enriquecimiento del perfil, como podemos observar la gran mayoría de las palabras del perfil de usuario utilizando **TF*IDF** (Figura 3.2) pueden ser englobadas con la palabra *computer science* que se muestran en el perfil enriquecido. Palabras como *language, program, programming, programme* y *object-oriented*, nos dan la información del interés del usuario por los lenguajes de programación. Sin embargo, palabras como *armed* y *military* no pertenecen a intereses en el repositorio del usuario y ni siquiera aparecen en el repositorio original. Los resultados a detalle de la nube del perfil enriquecido serán comentados en el Capítulo 5.

3.5 Resumen

Durante el desarrollo de este capítulo hemos visto de forma general el proceso que desarrollamos durante esta investigación en un marco formal de trabajo. Hablamos de la construcción del perfil y cómo la normalización de las palabras mediante el **TFIDF** nos permite obtener palabras más relevantes del repositorio de usuarios, esto lo pudimos observar a través de la visualizacion del perfil. La visualización del perfil nos permitió observar los pesos de las palabras mediante los tamaños que toman debido a su importancia dentro del repositorio, también nos dimos cuenta cómo los repositorios pequeños tienen palabras uniformes en tamaño. El mecanismo de enriquecimiento del perfil nos permitió observar cómo estos repositorios se volvían visualmente más atractivos.

Capítulo 4

Caso de estudio - Marcadores de Internet

Esta sección ha sido desarrollada para presentar el caso de estudio y la aplicación del modelo a este. Como primer punto hablaremos de los marcadores de Internet del usuario como fuente de información implícita. Como segundo punto, hablaremos de la obtención del repositorio: el método que llevamos a cabo para poder recopilar las páginas web a partir de los marcadores, la limpieza del material para procesarlo y la extracción del texto para su análisis.

4.1 Los marcadores de Internet como fuente de información implícita

El crecimiento que ha estado teniendo Internet tan acelerado y desmedido ha dejado a los usuarios en condiciones complejas para poder realizar una adecuada búsqueda de información así como su almacenamiento para una mejor localización posteriormente. Herramientas como los marcadores de Internet, agentes de búsqueda (*e.g.* www.trap.it, www.stumbleupon.com/ y www.google.com.mx/alerts) y sistemas de recomendación (*e.g.* www.coquetame.com/ y foursquare.com/about/) enfrentan estos escenarios pero necesitan de la colaboración del usuario para proporcionar información que deben utilizar para su mejor desempeño.

En la Figura 4.1 mostramos algunas fuentes de información que el usuario deja en su interacción con Internet.

Figura 4.1: Algunas fuentes de información del usuario.

En nuestro trabajo de investigación utilizaremos los marcadores de Internet guardados por los usuarios ya que la naturaleza de la acción de guardar una página web como un marcador, nos indica que el usuario tiene un interés sobre el contenido de esa página web. El almacenamiento por parte del usuario de sus marcadores de Internet, convierte a estos mismos en un repositorio de información personal el cual puede ser analizado para conocer estos intereses.

Hoy en día podemos manejar nuestros marcadores de Internet entre diversos dispositivos obteniendo mayor provecho de nuestra información almacenada. Entre los dispositivos como *teléfonos inteligentes, tabletas, computadoras portátiles y de escritorio* podemos acceder y guardar nuestras páginas web. Todo esto se puede llevar a cabo a través de un servicio como los que ofrecen algunas compañías de navegadores

web como *Mozilla Firefox*[1], *Opera*[2], *Google Chrome*[3], etc. que a través de una cuenta permiten sincronizar la información guardada, configuraciones, historial, etc, en un servicio de almacenamiento en la nube, de tal forma que los marcadores de Internet son una fuente con fácil acceso y menos invasiva para el usuario.

A través de la *minería de texto* nosotros queremos descubrir los intereses del usuario que se encuentran de forma implícita en los marcadores de Internet, para poder crear una representación visual de su perfil. Pero antes es importante, identificada la fuente de extracción de datos, obtener el repositorio para llevar a cabo las tareas de minería.

4.2 Obtención del repositorio

Tenemos a los marcadores de Internet como fuente de datos, de los cuales vamos a extraer la información del usuario que yace de manera implícita. Los marcadores de Internet son enlaces a páginas web que debemos obtener para hacer el repositorio de usuario como un grupo de documentos a procesar. Una vez que logremos conseguir los documentos web, debemos de hacer la extracción del texto para empezar el proceso de limpieza y dicriminar palabras que no representan información.

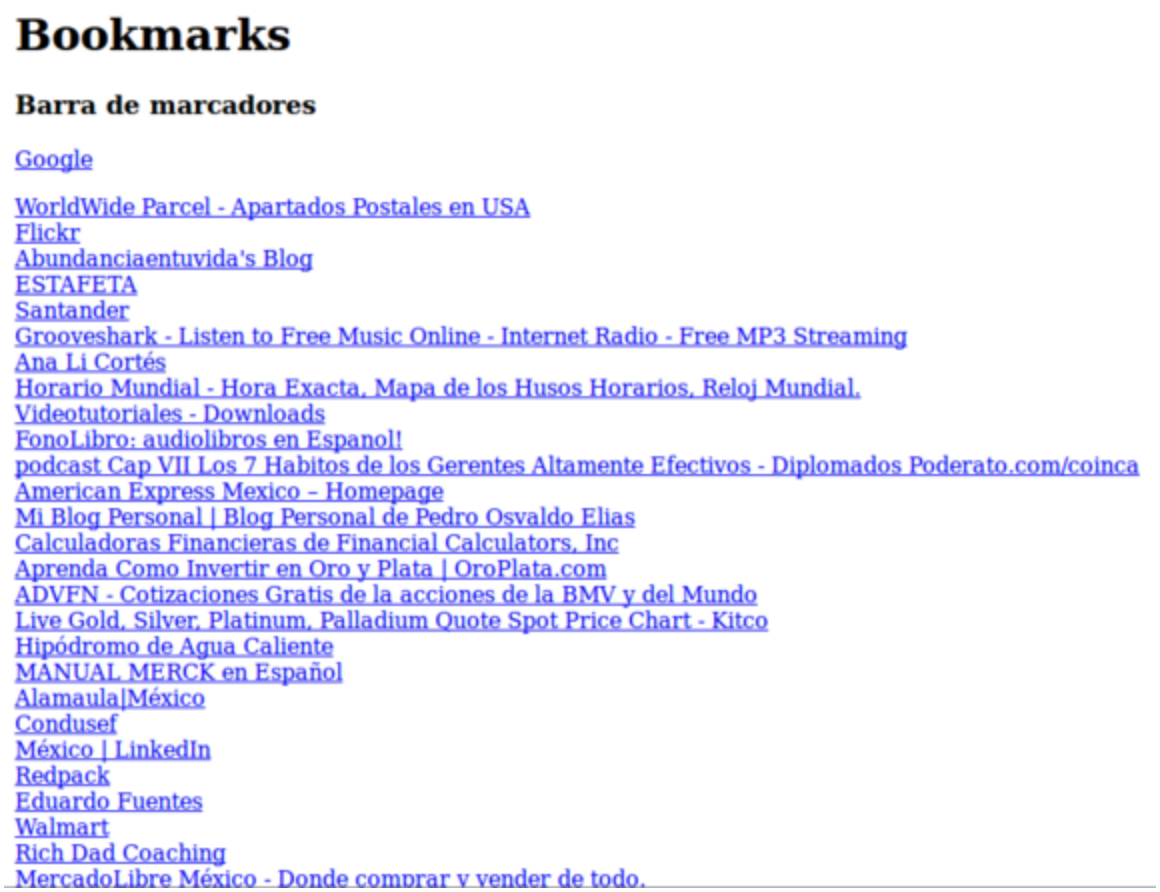
En las siguientes secciones daremos más detalles de estos tres procesos: obtención del repositorio, extracción del texto y limpieza.

4.2.1 Obtención de las páginas web a partir de los marcadores

Los marcadores de Internet son guardados en los navegadores web (*e.g.* Firefox, Chrome o IE) como enlaces a las páginas seleccionadas por los usuarios. Podemos obtener las páginas web a través del documento de exportación que generan los

[1] http://tinyurl.com/d35mgk6
[2] http://www.opera.com/link/
[3] http://tinyurl.com/alq8mso

navegadores web; el mecanismo de exportación genera un documento en formato **HTML** con los enlaces de las páginas web (Figura 4.2). Si el usuario ha creado carpetas para administrar sus enlaces, este mecanismo de exportación también coloca el nombre de dichas carpetas así como carpetas que contienen los navegadores web integradas desde su instalación.



Bookmarks

Barra de marcadores

Google

WorldWide Parcel - Apartados Postales en USA
Flickr
Abundanciaentuvida's Blog
ESTAFETA
Santander
Grooveshark - Listen to Free Music Online - Internet Radio - Free MP3 Streaming
Ana Li Cortés
Horario Mundial - Hora Exacta, Mapa de los Husos Horarios, Reloj Mundial.
Videotutoriales - Downloads
FonoLibro: audiolibros en Espanol!
podcast Cap VII Los 7 Habitos de los Gerentes Altamente Efectivos - Diplomados Poderato.com/coinca
American Express Mexico – Homepage
Mi Blog Personal | Blog Personal de Pedro Osvaldo Elias
Calculadoras Financieras de Financial Calculators, Inc
Aprenda Como Invertir en Oro y Plata | OroPlata.com
ADVFN - Cotizaciones Gratis de la acciones de la BMV y del Mundo
Live Gold, Silver, Platinum, Palladium Quote Spot Price Chart - Kitco
Hipódromo de Agua Caliente
MANUAL MERCK en Español
Alamaula|México
Condusef
México | LinkedIn
Redpack
Eduardo Fuentes
Walmart
Rich Dad Coaching
MercadoLibre México - Donde comprar y vender de todo.

Figura 4.2: Muestra parcial de un documento de exportación de marcadores.

Necesitamos obtener las páginas web de los enlaces del documento de exportación, así que desarrollamos un *script* para extraer las páginas web de los enlaces del documento de exportación.

Existen algunas páginas web con características singulares que fueron descartadas para esta investigación, como son:

Páginas guardadas con acceso restringido.

Algunos usuarios necesitan estar dados de alta en algunas páginas web para poder utilizarlas, como los foros o canales de información. Si los usuarios guardan estas páginas web como marcadores estando registrados, nosotros no podemos tener acceso a esas páginas ya que nos pedirían el nombre de usuario y contraseña para ver el contenido que guardó el usuario. Así que este tipo de páginas no pueden ser utilizadas en el análisis y deben de ser discriminadas.

Páginas que son accesos a archivos.

Hay páginas que son guardadas por ser un acceso a documentos de tipo *PDF*, documentos de *Office*, libros electrónicos, archivos comprimidos, librerías de lenguajes, etc. Estas páginas deben de ser tratadas de forma diferente para poder ser procesadas y sus archivos deben de ser identificados para ser leídos y poderlos tomar en cuenta para un análisis. En esta investigación, estos marcadores serán discriminados debido a que sólo estamos considerando el análisis de documentos web.

Páginas desarrolladas completamente en flash.

Existen páginas que contienen actividad multimedia realizada en *Flash*[4] como animación, películas, videojuegos, etc. En este tipo de páginas no se puede leer el interior debido a la naturaleza de compresión en formato binario de los archivos, de tal forma que para el análisis de esta investigación todas las páginas en los marcadores de los usuarios con esta naturaleza fueron descartadas.

Páginas en que todo el contenido es imagen.

Para darle un aspecto atractivo al sitio web, algunas páginas tienen todo su contenido en imagen, incluso el texto. Estas imágenes que contienen texto pueden ser procesadas con un mecanismo llamado *Reconocimiento Óptico de Caracteres* (**ROC** o en inglés **OCR** *Optical Character Recognition*), pero debido a que su población al interior de los repositorios procesados era muy pequeña no fueron utilizadas en esta investigación.

[4]http://www.adobe.com/es/products/flashplayer.html

Marcadores de los navegadores web.

Existen marcadores precargados para que el usuario acceda a páginas de las compañías de los navegadores web, estas páginas contienen información de sus productos, la promoción de canales de información, accesos a noticias, foros, etc. Si no removemos estos marcadores del análisis, aparecerían en los resultados de su perfil y parecería que el usuario tiene estos intereses.

4.2.2 Extracción de texto

Las páginas web extraídas contienen texto al interior de etiquetas de **HTML**. El **HTML** (*Hyper Text Markup Language* o *Lenguaje de Marcado de Hipertexto*) es la estructura (Figura 4.3) que a través de etiquetas de inicio y de cierre que se encuentran entre corchetes angulares (< >) forman los contenidos de las páginas web. Por ejemplo, una etiqueta de inicio como <a> es utilizada para la inserción de enlaces en el documento web y se coloca una etiqueta de cierre como para indicar que ahí termina.

```
<HTML>
   <META>
   <HEAD>
     <TITLE>Título de la página</TITLE>
   </HEAD>
   <BODY>
      Aquí iría el contenido de la página
   </BODY>
</HTML>
```

Figura 4.3: Estructura de HTML básica para generar una página web.

Las etiquetas de **HTML** que nos importan para nuestra investigación, son aquellas que tienen una tarea directamente con el texto del documento. Las etiquetas que decidimos seleccionar son aquellas que dan alguna relevancia en *la estructura del documento*, aquellas que dan características en *el tamaño de texto, color, tipo de letra* y las que enfatizan el texto como *tachado, subrayado, negrita* o *cursiva*.

Dos etiquetas que tienen relevancia en la estructura de los documentos web son las etiquetas de creación de párrafos `<p> </p>` y la creación del título de la página web `<title> </title>`, estas nos dan palabras claves de las que podemos extraer información relevante del contenido del documento. Asumimos también que, si los desarrolladores de estos documentos web han creado énfasis con algunas de estas etiquetas sobre algún texto, esto nos indica que dicho texto contiene una relevancia en el tema del documento.

La naturaleza de la etiqueta `<meta>` nos indica que debemos tomarla en cuenta para esta investigación debido a que es utilizada para colocar una descripción breve de la página web y es utilizada también para colocar palabras claves o tópicos relevantes del contenido, pero debido a su mal uso creemos que contiene más elementos de ruido que elementos que puedan contribuir con nuestro objetivo. La etiqueta `<meta>` se utilizó por muchos buscadores web para el posicionamiento en los resultados de las búsquedas web; debido a esto, los desarrolladores empezaron a colocar palabras al interior de esta etiqueta que eran populares en Internet pero las páginas web no contenían nada relacionado en su interior con esos temas, así que sólo eran colocadas para poder mejorar el posicionamiento en los resultados de las búsquedas. Compañías de Internet como *Google* no toman en cuenta[5] el contenido de esta etiqueta debido a este mal comportamiento de los desarrolladores.

Las etiquetas que fueron seleccionadas (Tabla 4.1) son de la especificación 4.01 del **HTML** y a pesar de que algunas de ellas estén marcadas como obsoletas son utilizadas todavía por los desarrolladores y aún son interpretadas por los navegadores web. También todas las páginas anteriores a la especificación utilizan esas etiquetas.

Las reglas de la especificación 4.01 sobre el uso de las etiquetas a pesar de haber sido creadas como un estándar, no son seguidas por toda la comunidad de los desarrolladores web y existen documentos web mal estructurados y con mal uso de estas etiquetas, a pesar de esto, los navegadores web las interpretan y muestran su contenido.

[5]http://tinyurl.com/a27uemm

Etiquetas	Descripción
<a>	*Define un enlace*
<b>	*Define texto en estilo Bold*
<big>	*Define texto en tamaño grandes*
<basefont>	*Establece un color, tamaño y letra por defecto. OBSOLETO*
	Centra un texto. OBSOLETO
	Enfatiza el texto
<font>	*Define letra, color y tamaño.* OBSOLETO
<h1>	*Define título de texto en forma 1*
<h2>	*Define título de texto en forma 2*
<h3>	*Define título de texto en forma 3*
<h4>	*Define título de texto en forma 4*
<h5>	*Define título de texto en forma 5*
<h6>	*Define título de texto en forma 6*
<i>	*Define texto en itélica*
<p>	*Define los pérrafos del texto*
<pre>	*Define texto con preformato*
<s>	*Define texto tachado.* OBSOLETO
<small>	*Define texto en tamaño pequeño*
<strike>	*Define texto tachado.* OBSOLETO
	Define texto en negrita
<title>	*Define el título de la página web*
<u>	*Define texto subrayado.* OBSOLETO

Tabla 4.1: Etiquetas de HTML consideradas para la extracción de contenido.

Durante el desarrollo de esta investigación se encontraron algunas páginas que tenían un muy mal uso de las etiquetas que vale la pena mencionar como observación, por ejemplo:

- Una página web que utilizaba la etiqueta `<body>` como indicador de párrafos.
- Etiquetas de listas para la creación del menú de la página en lugar de la etiqueta `<menu>`.
- Etiquetas de tabla para crear toda la estructura del documento en lugar de etiquetas `<div>`.
- Texto al interior de la etiqueta `<body>` sin utilizar etiquetas `<p> </p>` para darle estructura.
- Páginas sin etiqueta `<title>` o `<body>`.

Debido a que nuestra propuesta esta basada en las reglas del **HTML** 4.01, todo el texto que se encuentre en etiquetas fuera de nuestra lista, será descartado en el proceso de limpieza del repositorio de los usuarios.

El primer paso que debemos hacer es tomar las etiquetas que nos interesan del **HTML**. Para hacer la selección de aquellas que son de nuestro interés y discriminar el resto, creamos un programa que nos genera una estructura de tipo árbol de las etiquetas del documento. Debido al comportamiento de árbol que lleva a cabo la herramienta, si una etiqueta es un nodo padre y no es tomada en cuenta, sus hijos tampoco aparecerán. Este fue el primer obstáculo que encontramos debido a que la etiqueta `<div>` y `<span>` son etiquetas con una función de contenedores al interior del documento web y al momento de no ser tomadas en cuenta, se elimina mucho contenido importante que se encuentra dentro de etiquetas que sí nos interesan. Así que en la aplicación, fueron tomadas en cuenta sólo para poder acceder a los elementos que contenían. Se desarrolló una aplicación que imprime a un documento de texto las etiquetas seleccionadas para poder apreciar su contenido.

Otro caso que es bueno mencionar, es el de la etiqueta `<body>` que contiene todo el cuerpo del contenido que será mostrado por los navegadores; esta etiqueta, si no la tomamos en cuenta, no imprime nada al documento debido que la etiqueta `<body>` funciona como padre de todo el contenido del documento web. Su comportamiento es como el de una raíz así que la etiqueta `<body>` siempre debe ser tomada en cuenta en este mecanismo.

La aplicación genera un documento de texto (Tabla 4.2) con el nombre del documento web y al interior del documento de texto genera el nombre de la etiqueta al inicio de cada linea de texto sin los corchetes angulares (`< >`) y brincando a una linea nueva al empezar una nueva etiqueta, de esta manera podemos visualizar lo que contiene cada etiqueta de forma más clara y traer el texto que nos interesa para procesar cada documento.

title Casos de éxito: Google Apps for Education
a Apps for Education
a Acceder
div www.
div y
strong Soluciones
a Google Apps (gratis)
a Google Apps for Business
a Google Apps for Education
a Google Apps for Government
a Comparar ediciones
a Convertirse en distribuidor
strong Productos
a Gmail

Tabla 4.2: Ejemplo parcial de contenido del texto generado a través de la aplicación.

4.2.3 Limpieza

Existen algunas observaciones que vale la pena mencionar si vamos a intervenir el texto de una página web. Los puntos que vamos a mencionar fueron el resultado de la experimentación realizada con los repositorios de los usuarios con los que interactuamos, los puntos son los siguientes:

Contenidos en diferentes idiomas.

Las páginas web de los usuarios analizados en su mayoría tenían contenido en español y en inglés, pero existían algunas páginas en otros idiomas, aunque pocas. Para poder tratar estos casos es necesario poder tener todo el material en un solo idioma. Un ejemplo de complicación en este tema, es una página web con contenido de películas y sus títulos en los idiomas originales.

Semántica de la lengua y errores.

Dependiendo del idioma, cada regla es diferente. En los repositorios manejados encontramos la mayor parte de las páginas en español. Las palabras correctamente escritas con acento y sin acento, palabras acentuadas de manera incorrecta y palabras mal escritas representan cuatro grupos de palabras diferentes. Pero si decidimos tomar el idioma inglés como nuestro idioma base para el análisis, solo tendríamos dos grupos de palabras: las correctamente escritas y las mal escritas.

Distorsiones de la lengua escrita.

Durante esta investigación en los documentos de HTML nos encontramos con un fenómeno propio del comportamiento de los usuarios en Internet, y es el de una escritura diferente a la normal. Este fenómeno de la escritura en Internet tiene cuatro comportamientos que pudimos observar en el repositorio analizado:

1 Sustituir números por letras para hacer las palabras, por ejemplo: *enlace* por *3n14c3* o *gasto* por *64570*. En la Tabla 4.3 podemos ver la lista

completa de sustituciones de números por letras.

II Usar acrónimos en lugar de frases, por ejemplo: *lol* por *lot of laugh* (mucha risa) o *FYI* por *for your information* (para su información).

III Escribir con letras mayúsculas, minúsculas y signos de puntuación, por ejemplo: *princesa* por `--Pr1NceS@--` o *campeón* por `!!!!C@mP3On!!!!`

IV Escribir las palabras juntas, por ejemplo: *CursoTutoriales* o *VideoTutoriales*.

V El regionalismo de la lengua, un comportamiento que se da en diferentes áreas geográficas de un país. Por ejemplo; *apa* que con acentuación o si ella significa papá o padre en el área del norte de México y no *apa* que significa *American Psycological Association*[6]; o *camara*, que es una expresión que se utiliza en el centro de México y que no significa *cámara fotográfica*.

El número de marcadores guardados por los usuarios.

Si tenemos en cuenta la discriminación de los puntos anteriores y a eso le agregamos un tamaño no muy grande de marcadores guardados por el usuario, empezamos a tener problemas con la dimensión del repositorio que podemos utilizar para hacer nuestro análisis. Pero para fines de las hipótesis marcadas en esta investigación todavía podemos utilizar un grupo pequeño de páginas web además los resultados serán evaluados por los usuarios al final.

Las observaciones marcadas en los puntos anteriores resultan un reto muy importante para la minería de texto y es un tema de investigación aparte del que nosotros tratamos aquí. Los puntos *Distorsión de la lengua escrita* así como *Semántica de la lengua y errores* los podemos encontrar en los foros o los comentarios que la gente hace en las páginas web. El último pero no menos importante, *El número de marcadores guardados por los usuarios* es uno de los puntos más complicados debido a que los trabajos de minería son más enriquecedores cuando se trabaja con grandes cantidades de datos a procesar. En este trabajo proponemos un mecanismo

[6] www.apa.org/

Números a utilizar	**Letras a sustituir**
0	O
1	L
2	Z
3	E
4	A
5	S
6	G
7	T
8	B
9	P

Tabla 4.3: Números utilizados para sustituir letras.

que puede hacer frente a este escenario y en el punto del enriquecimiento del perfil ahondaremos más en el tema.

Teniendo generados los documentos, encontramos que los caracteres con acentuaciones no aparecen correctamente y en su lugar están códigos en HTML tanto en nombre como en número. Buscamos en la especificación del HTML 4.01 la lista completa para poder completar las palabras correctamente. La lista completa utilizada durante esta investigación se muestra en la Figura 4.4.

La principal característica de este fenómeno observado es que las acentuaciones de las palabras no salían y eso causaba que las palabras estuvieran incompletas; pero con la lista de los códigos correctos de HTML, logramos cambiar de los documentos de texto todos estos códigos por las acentuaciones correctas así como las letras $\tilde{n}$ $\tilde{N}$ faltantes. Todos aquellos códigos de HTML que representaban símbolos como el más (+), menos (-), los de interrogación (¿?), etc., no nos interesan y fueron eliminados en el proceso de limpieza de los documentos.

Símbolo	Número en HTML	Nombre en HTML	UTF-8
á	á	á	Ã¡
é	é	é	Ã©
í	í	í	Ã
ó	ó	ó	Ã³
ú	ú	ú	Ãº
ñ	ñ	ñ	Ã±
Á	Á	Á	Ã·
É	É	É	Ã‰
Í	Í	Í	Ã
Ó	Ó	Ó	Ã“
Ú	Ú	Ú	Ãš
Ñ	Ñ	Ñ	Ã‘

Figura 4.4: Caracteres a sustituir.

Utilizamos herramientas de **UNIX**[7] para ayudarnos en la limpieza de los documentos para llevar a cabo la sustitución y eliminación de caracteres, códigos y palabras que no necesitamos para el procesamiento del repositorio. A través de una serie de pequeños programas desarrollados en *shell* fuimos haciendo el proceso de limpieza.

La siguiente lista representa el orden de la limpieza que llevamos a cabo en los repositorios de los usuarios.

1. Eliminamos del documento las filas que empiezan con las etiquetas de HTML que no nos interesan.

2. Eliminamos el inicio de cada fila (que es el indicador de las etiquetas que deseamos analizar).

3. Buscar y remplazar los códigos **HTML** o **UTF-8**[8] por las acentuaciones co-

[7] http://www.unix.org/

[8] 8-bit Unicode Transformation Format

rrectas.

4. Eliminamos signos de puntuación(Figura 4.5).

5. Eliminamos números.

6. Eliminamos meses y días de la semana.

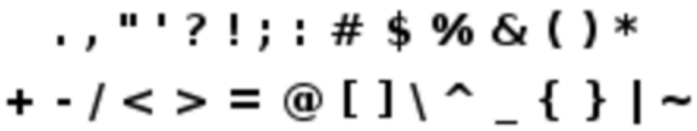


Figura 4.5: Signos de puntuación eliminados.

Casos de éxito Google Apps for Education
Apps for Education
Acceder
Soluciones
Google Apps gratis
Google Apps for Business
Google Apps for Education
Google Apps for Government
Comparar ediciones
Convertirse en distribuidor
Productos
Gmail

Tabla 4.4: Ejemplo parcial de contenido del texto ya procesado por el primer mecanismo de limpieza.

Como podemos observar en la Tabla 4.4, palabras en inglés y español están presentes en el documento. Como mencionamos anteriormente, habíamos escogido el idioma inglés como el idioma para trabajar el repositorio de los usuarios, así que nos

dimos a la tarea de pasar al inglés los documentos que estaban en español y poder continuar.

Una vez que todos los documentos los teníamos en el idioma inglés, pasamos a la segunda etapa del proceso de limpieza que consiste en dos puntos:

- Eliminación de la palabras vacías.
- Eliminación de las palabras comunes de los manejadores de contenidos.

Durante el proceso de limpieza llevamos a cabo la eliminación de *palabras vacías* (el término en inglés es **stopword**) de los documentos de los usuarios. Estas palabras se les llama así debido a que no representan información y se encuentran mucho en la mayoría de los documentos. Si utilizamos un proceso de análisis de texto en donde tomemos en cuenta la frecuencia de las palabras, este grupo de palabras serían las que tomarían los primeros lugares.

No existe una lista oficial de estas palabras y nosotros utilizamos una lista que creamos de fuentes[9] [10] diferentes que localizamos en Internet, la lista completa utilizada en esta investigación la mostramos en el Anexo A.

all	do	etc	that	well
almost	does	even	thats	went
alone	doesn	ever	the	were
along	doesnt	every	their	weve
alongside	doing	everybody	theirs	what
already	done	everyone	them	whatever
also	dont	everything	themselves	whatsoever
although	down	everywhere	than	when

Tabla 4.5: Lista breve de palabras vacías.

[9] http://tinyurl.com/yz83wo6

[10] http://www.ranks.nl/resources/stopwords.html

Las palabras vacías no son el único fenómeno que puede intervenir debido a su mayoría al interior de un documento, en Internet existe un caso más que debemos considerar que tiene un comportamiento similar en repetición, pero no son necesariamente palabras vacías.

En Internet se hace cada vez más grande el uso de manejadores de contenido **CMS** (*Content Managment System*), este tipo de sistemas son utilizados para generar páginas web y administrarlas de manera más sencilla. Un caso que podemos mencionar como ejemplo en donde se implemente un manejador de contenidos sería un portal web de noticias que maneja mucha información y publica durante todo el día. La mejor herramienta que un portal así requiere, es un manejador de contenido en donde el administrador da de alta a los usuarios que van a publicar; y los usuarios que publican, sólo se preocupan por escribir las notas sin pensar en nada referente a la informática. *Drupal*[11], *Wordpress*[12] y *Joomla*[13] son algunos de los manejadores de contenido más utilizados en Internet hoy en día, pero el uso de este tipo de administradores agrega no sólo una forma correcta en la estructura de HTML de la página web sino también una serie de palabras que se repiten en las páginas generadas por los usuarios debido a que son plantillas que se definen para la creación de contenido. Creemos conveniente que este listado de palabras que los generadores de contenido contienen sean discriminadas del análisis para que no influyan en el texto a analizar.

Podemos mencionar un escenario en donde un usuario tenga muchos marcadores de Internet que utilicen *Wordpress*, pueden ser portales de noticias de diversas áreas de interés, portales como blogs o foros. Si hiciéramos un procesamiento de estas paginas web nos saldría en los resultados la palabra *Wordpress* como si fuera uno de los intereses que tiene el usuario, y no es así, debido a que esa palabra aparece en los resultados no por un interés del usuario sino por su peso en el repositorio.

[11] drupal.org
[12] wordpress.com
[13] joomla.org

Debido a esta observación se generaron dos listas a discriminar durante esta investigación, una con palabras sencillas y otras con palabras dobles o más. Las listas fueron generadas previamente de unos agrupamientos de palabras que fueron localizadas en diversas secciones de los **CMS** de los repositorios con los cuales se interactuó para este experimento. Los nombres de estas agrupaciones los proponemos debido a su propósito.

La agrupación de *Publicación* (Tabla 4.6), son enlaces que tienen los **CMS** para que a través de estos medios puedan interactuar con el web site; por ejemplo, si el usuario esta registrado en *Twitter*[14] puede dejar comentarios con su cuenta en el sitio web en donde esté, así mismo con su cuenta de *Facebook*[15], *Google*[16], *Del.icio.us*[17], etc., o hacer publicaciones de los contenidos de la página web como enlaces a través de sus cuentas de redes sociales para que sus amistades puedan visitar también estos contenidos que gustaron al usuario, de tal forma que estos enlaces son comunes debido a que en Internet es necesario tomar un comportamiento de divulgación.

flickr
google
twitter
facebook
del.icio.us

Tabla 4.6: Enlaces encontrados en CMS para interactuar con el contenido.

El grupo de *CMS* (Tabla 4.7) contiene enlaces que podemos encontrar en la mayoría de las ocasiones en la base de las páginas de los manejadores de contenido y son accesos que se dirigen al sitio web de los manejadores de contenido para ayudar a difundir su uso en Internet. No importa en que sección de la página web nos encontremos estos enlaces aparecen en la plantilla y son muy repetitivos independientemente de los contenidos de las páginas.

[14] https://twitter.com/
[15] www.facebook.com/
[16] www.google.com
[17] www.delicious.com

wordpress
drupal
blogger
joomla
creative commons

Tabla 4.7: Enlaces que los CMS manejan como base en sus plantillas.

Opciones del sitio web (Tabla 4.8), son opciones que vienen en el menú o el submenú del sitio web. Con enlaces a las secciones que contiene el sitio web para que el usuario pueda contar con accesos rápidos a estas áreas; por ejemplo: la sección de *contacto* o la sección de *acerca de nosotros* o *quiénes somos*.

Registro del sitio (Tabla 4.9), es una agrupación con las palabras relacionadas al registro de usuarios en los sitios web que manejan este requisito para la interacción con los usuarios, contenidos publicados, manejo de accesos a secciones especiales o de algún servicio.

La agrupación de *Comunidad* (Tabla 4.10) es una sección en donde encontramos enlaces a secciones en donde el sitio de Internet busca tener un acercamiento con los usuario visitantes para formar una comunidad como por ejemplo, los foros.

Interacción del usuario (Tabla 4.11) es una agrupación de palabras que suelen ser enlaces a acciones a tomar con el contenido que la página web publica. Por ejemplo, en un artículo publicado el usuario puede comentar, puede votar que tan bueno es, puede ver la calificación que los demás le han dado, puede imprimirlo o dar clic para poder ver más del artículo, etc.

La agrupación de *Sin clasificar* (Tabla 4.12), son palabras que no fueron agregadas a una agrupación de la mencionadas anteriormente ya que se encontraban en varias secciones.

Consideramos que en un trabajo futuro se pueda ampliar estas agrupaciones de palabras para proyectos de minería de texto en trabajos relacionados con documentos

privacy and security	help	terms
conditions	privacy	policy
select a language	sitemap	faqs
assistence and support	faq	search
support	powered	copyrights
email	online	suggestions
press and blogs	directory	files
archive	press release	about us
press and communication	my cart	site map
news and blogs	most popular	copyright
rights reserved	terms of use	contact us
terms and conditions	main menu	social media
home	privacy practices	search tools
legal notices	terms of service	send an email
privacy notice	privacy policy	join the team
contact form	social media	quality policy
online payment	change country	other countries
language support	search people	about
privacy and conditions	report a bug	feedback
enhance your experience	website	mobile version

Tabla 4.8: Enlaces que los CMS manejan en menús y/o sub menús.

initiation	registration
register	registry
session	lost password
start	account
enter	subscribe
login	members
join	forgot your password
i forgot my password	logout
password recovery	sign in
sign up	user login
log in	log out
my account	new user

Tabla 4.9: Enlaces que los CMS manejan referente al registro de usuarios.

rss
feed
community
groups
categories
forums
forum
blogs
blog

Tabla 4.10: Enlaces que los CMS manejan para tener un acercamiento con los usuarios.

comments	comment	cite this page
content	followers	follow
posted	posts	post
entries	recent	shared
share this page	vote	click
popular	ratings	rate
favorites	friend	now
tags	print version	see more
see less	see all	follow up
follow us	more sites	more news
more articles	print this page	share
email this page	upload a file	cite
chat online	live chat	most viewed

Tabla 4.11: Enlaces que los CMS manejan para la interacción con el usuario.

videos	video
music	advertising
advertise	ads
buy	sample
pdf	download
images	web links
click here	create a page

Tabla 4.12: Lista de enlaces no clasificados.

o páginas web, no sólo hacer un trabajo que enriquezca la lista de palabras sino de manejadores de contenido analizados.

4.3 Resumen

Durante este capítulo hemos visto el proceso de obtención de los repositorios con los cuales se trabajó durante esta investigación, vimos las diferentes características que tienen las páginas web y los retos que cada característica presenta para la extracción de datos. Realizamos varias observaciones necesarias antes del proceso de limpieza y las describimos cada una de ellas, para lograr obtener un repositorio personal como *saco de palabras*, previo al procesamiento a la construcción del perfil.

Capítulo 5

Experimentación, pruebas y resultados

Durante este capítulo expondremos la experimentación realizada con repositorios de información personal a una muestra de 30 individuos. Se expondrán las pruebas de las visualizaciones así como los detalles de la evaluación que se llevó a cabo con los usuarios, observaremos los resultados y mostraremos si los resultados obtenidos pueden ayudarnos a relacionar a los usuarios. El objetivo de las pruebas es ver si los perfiles extraídos efectivamente representan los intereses de los usuarios. Debido a esto, se les presentaron varias visualizaciones de perfiles a los usuarios para que los calificaran de acuerdo a qué tanto reflejaban temas de su interés. También se aprovechó el hecho de que la mayoría de los usuarios contactados para las pruebas forman un grupo en la vida real (a esto se le llama *ground truth*) y este grupo comparte intereses. En este sentido, debería ser posible recuperar dicho grupo basándose en los perfiles de usuario. Para ello, se tomaron los perfiles como documentos de texto y se obtuvo una matriz de similitud. Aunque los experimentos no son exhaustivos, los resultados obtenidos en ambas pruebas confirman la hipótesis (el procesamiento de los repositorios arroja perfiles que representan intereses del usuario).

5.1 Creación de perfiles de usuario

El proceso de creación de los perfiles de usuario nace de la obtención y depuración de datos del usuario para su transformación en información. Ambos procesos fueron detallados durante el Capítulo 4, en donde hablamos de cómo elegimos la fuente para extracción de información e hicimos algunas observaciones de escenarios encontrados, así como de características propias de Internet y el reto que representan. En el Capítulo 3, se detalló la creación de los perfiles de usuario y cómo aportamos un mecanismo de enriquecimiento para la información extraída.

5.1.1 Muestra de usuarios y repositorios

La muestra (Tabla 5.1) de usuarios que colaboró en esta investigación fue de 30 personas todos mayores de edad y, en su mayoría, estudiantes de licenciatura y posgrado de la Facultad de Ingeniería Mecánica y Eléctrica; de estos, 18 eran del área de licenciatura y ocho del área de posgrado. La muestra restante está conformada por dos mujeres y dos hombres que forman parte de matrimonios jóvenes.

Sector	**Muestra**	**Edades**
Licenciatura	18	18 - 20
Posgrado(Maestría)	8	24 - 29
Restante	4	28 - 34

Tabla 5.1: Representación de la muestra.

Durante el desarrollo de este capítulo, marcaremos a los usuarios con una U posteriomente el número que lo identifica *e.g.* U 25. Como forma de identificación hemos utilizado un número para cada usuario que no tiene ningún ordenamiento en especial, y principalmente, para guardar las identidades de los usuarios debido a que este punto es un tema delicado. Es precisamente por este tema que la muestra utilizada en esta investigación no fue mayor. Incluimos en la Tabla 5.3 algo de estadística descriptiva de la muestra así como también colocamos en La Tabla 5.2 el número de

documentos web que fueron procesados por usuario en nuestra muestra.

U 1	16	**U 16**	80
U 2	5	**U 17**	6
U 3	32	**U 18**	10
U 4	10	**U 19**	6
U 5	66	**U 20**	6
U 6	17	**U 21**	7
U 7	6	**U 22**	5
U 8	65	**U 23**	19
U 9	7	**U 24**	11
U 10	12	**U 25**	108
U 11	5	**U 26**	19
U 12	51	**U 27**	9
U 13	36	**U 28**	44
U 14	24	**U 29**	9
U 15	148	**U 30**	25

Tabla 5.2: Tamaño de los repositorios de los usuarios.

En la Tabla 5.4 nos muestra como apenas nueve de los usuarios contienen un número considerable de documentos a procesar, frente a un grupo de doce usuarios (Tabla 5.5) con repositorios menores. Es importante señalar que aunque sean pocos documentos web eso no significa que su contenido sea escaso, es decir, puede haber un documento muy extenso y que enriquezca nuestro proceso de creación de perfiles.

Separar estos dos grupos de muestras mayores (Tabla 5.4) y menores (Tabla 5.5) nos va a servir para observar su comportamiento en la evaluación y ver si podemos encontrar alguna característica. Además buscamos observar cómo evalúan los usuarios de repositorios menores, sus visualizaciones del perfil enriquecido.

Estadística	Valor
Media	28.8
Mediana	14
Mínimo	5
Máximo	148
Varianza	1170.16
Desviación estándar	34.2

Tabla 5.3: Estadística descriptiva de la muestra utilizada.

U 3	**U 5**	**U 8**
32	66	65
U 12	**U 13**	**U 15**
51	36	148
U 16	**U 25**	**U 28**
80	108	44

Tabla 5.4: Grupo de usuarios con mayor número de documentos a procesar.

U 2	**U 4**	**U 7**	**U 9**	**U 11**	**U 17**
5	10	6	7	5	6
U 19	**U 20**	**U 21**	**U 22**	**U 27**	**U 29**
6	6	7	5	9	9

Tabla 5.5: Grupo de usuarios con menor número de documentos a procesar.

5.1.2 Perfiles de usuario

Con los repositorios recolectados se obtuvieron los perfiles de los usuarios antes mencionados, bajo el proceso descrito en el capítulo del Modelo (3). En este punto se tiene un documento de texto con la suma (el *top 20* de cada documento) de las palabras significativas de todos los documentos que forman el repositorio de información personal. Con este repositorio ya formado como *el perfil de usuario*, generamos otro repositorio más: el enriquecido (el proceso de enriquecimiento del perfil es descrito en la sección 3). Conformados ambos perfiles de usuario podemos generar las visualizaciones a través de una herramienta, para apreciar los intereses del usuario y ver las diferencias entre ambas. La herramienta utilizada para esta investigación es **Wordle**[1]; con ella colocamos el archivo de texto del perfil del usuario y, bajo una selección de 150 palabras por perfil, obtenemos ambas visualizaciones para poder llevar a cabo las evaluaciones.

5.2 Calificación de perfiles de usuario

Figura 5.1: Imagen *Dummy* mostrada a los usuarios en la evaluación.

[1]http://www.wordle.net/

De los resultados de las evaluaciones esperamos que el *Dummy* (Figura 5.1)sea el menos votado, consideramos que el mecanismo de enriquecimiento propuesto en esta investigación generará votaciones por encima de las visualizaciones generadas con *TFIDF*, aún y que el mecanismo ha sido propuesto ante repositorios menores, creemos que puede ser útil ante repositorios de diversos tamaños. En la encuesta no sólo exponemos las visualizaciones de los perfiles enriquecidos y el generado mediante *TFIDF*, para observar cómo son evaluadas por los usuarios. También generamos una visualización de palabras sin normalizar para observar el comportamiento en las votaciones con respecto a la visualización generada mediante *TFIDF*, creemos que en la evaluación marcarían como cuarto sitio a esta visualización, por debajo de la generada con *TFIDF*. Y por último, la visualización *Dummy*, debido a que los resultados de los perfiles de los usuarios mostraron cero interés en temas como los mostrados en esta visualización; por lo tanto, creemos que las votaciones de los usuarios en la encuesta van a ser bajas para el *Dummy*.

5.2.1 Configuración

En la encuesta se utilizaron tres hojas tamaño carta, con la siguiente distribución:

- La hoja de inicio contiene las instrucciones y la explicación de la encuesta.
- La segunda hoja contiene la visualización *TFIDF* y debajo de ella la visualización del perfil enriquecido a través de *WordNet*.
- La tercera hoja presenta la visualización del perfil según la frecuencia de palabras y debajo la visualización *Dummy*.

Previo a las instrucciones de cómo evaluar las visualizaciones, se les explicó cómo las visualizaciones de nubes eran formadas por palabras que estaban relacionadas entre sí y que el tamaño de las palabras era un indicador de su relevancia entre las

demás. Se les indicó que evaluaran las visualizaciones indicando cuál de ellas contenía las palabras con las que más sentía interés; cabe resaltar que a los usuarios no se les mencionó que las visualizaciones que estaban evaluando tenían relación con ellos.

Se utilizó un mecanismo de evaluación (Figura 5.2) en donde los usuarios rellenan estrellas (Pang y Lee, 2005), para indicar el grado de interés que sienten con la representación (*nube*) donde *cinco estrellas* significa *Muy Bueno*, *cuatro estrellas Bueno*, *tres estrellas Regular*, *dos estrellas Malo* y *una estrella* significa *Muy Malo*.

Figura 5.2: Imagen con estrellas para evaluar por los usuarios.

5.2.2 Resultados

Los resultados de la muestra completa de usuarios puede ser encontrada en el Apéndice B.

En los resultados de la Tabla 5.6 es claro que los usuarios sienten que las visualizaciones de los perfiles con *TFIDF* tienen más palabras de su interés y esa misma tendencia la podemos señalar con la visualización por *frecuencia de palabras*. Sin embargo las visualizaciones generadas con el mecanismo de enriquecimiento no salieron muy indiferentes a los usuarios, muy probablemente debido a que encontraron

prominentes palabras que engloban muchos de sus intereses. Por otro lado la visualización *Dummy* salió como esperabamos, hasta el final de esta encuesta; sin embargo, hubo usuarios que dieron calificaciones altas a esta visualización, esto se debe a que los usuarios tienen intereses en esta visualización pero no eran intereses que se encontraran almacenados en sus repositorios. Esto nos da una evidencia de que no todos los intereses del usuario podrán ser encontrados bajo una sola fuente de información como la que utilizamos en esta investigación, creemos que es necesario seguir haciendo experimentos con más fuentes de información y ver su comportamiento.

Nube	**Estrellas**
TFIDF	4.52
Enriquecido	3.38
TF	3.79
Dummy	2.28

Tabla 5.6: Valor promedio extraído de los resultados de las evaluaciones.

En la Tabla 5.7 se muestran los resultados de las evaluaciones que los usuarios con mayor número de documentos hicieron. Como podemos observar los resultados para la visualización generada con *TFIDF* son altos, con sólo una votación de tres y otra de cuatro estrellas. Se puede apreciar que las votaciones para la visualización de *frecuencia de palabras* (*TF*) salen mejores que las visualizaciones del *perfil enriquecido*, que presenta calificaciones regulares y por de bajo de regular.

El *usuario 25* (U 25) en su evaluación puso una calificación máxima (de cinco estrellas) a la visualización generada con *TFIDF* y una calificación de tres a la visualización del *perfil enriquecido*, lo cual nos indica que en la visualización con *TFIDF* encontró más palabras de su interés. Por otra parte la visualización del *perfil enriquecido* con calificación de regular nos muestra cómo el usuario pudo sentir interés por palabras que predominan como, *science, computer, computing* y *system*, pero pudo haberse sentido indiferente con las palabras *military, armed* y *war*. Si observamos la calificación regular es una calificación media en donde los usuarios no tienen una tendencia firme ni positiva ni negativa.

Nube	**U 3**	**U 5**	**U 8**
TFIDF	5	5	4
Enriquecido	1	5	3
TF	3	4	2
Nube	**U 12**	**U 13**	**U 15**
TFIDF	5	5	3
Enriquecido	4	3	2
TF	5	4	4
Nube	**U 16**	**U 25**	**U 28**
TFIDF	5	5	5
Enriquecido	2	3	4
TF	5	4	3

Tabla 5.7: Resultados del grupo de usuarios con mayor número de documentos a procesar.

Si sacamos el promedio (Tabla 5.8) de las evaluaciones de este grupo de usuarios podemos observar que no son muy diferentes de los promedios de la muestra (Tabla 5.6).

Nube	**Estrellas**
TFIDF	4.6
Enriquecido	3
TF	3.77

Tabla 5.8: Valor promedio del grupo de usuarios con mayor número de documentos.

La Tabla 5.9 nos muestra las evaluaciones de las visualizaciones de los perfiles de los usuarios con menor cantidad de documentos web procesados y cómo podemos ver los resultados se mantienen, teniendo las mejores votaciones las visualizaciones generadas con *TFIDF* seguidas de las realizadas con *frecuencia de palabras*. En esta tabla también podemos observar que las visualizaciones realizadas con el mecanismo de enriquecimiento obtienen votaciones promedio y bajas como mayoría.

Nube	**U 2**	**U 4**	**U 7**	**U 9**	**U 11**	**U 17**
TFIDF	5	4	5	4	5	5
Enriquecido	4	3	5	2	3	3
TF	4	2	4	5	3	3
Nube	**U 19**	**U 20**	**U 21**	**U 22**	**U 27**	**U 29**
TFIDF	5	3	5	5	5	2
Enriquecido	3	5	1	4	2	5
TF	5	5	5	4	3	4

Tabla 5.9: Resultados del grupo de usuarios con menor número de documento.

Nube	**Estrellas**
TFIDF	4.416
Enriquecido	3.33
TF	3.916

Tabla 5.10: Valor promedio del grupo de usuarios con menor número de documento.

Los valores promedios (Tabla 5.10) de las votaciones (Tabla 5.9) de este grupo de usuarios con repositorios menores nos muestra cómo el valor en la preferencia de los perfiles generados con *TFIDF* se mantiene al igual que las visualizaciones de los *perfiles enriquecidos*. Sin embargo las visualizaciones de *frecuencia de palabras* salen más altas, esto creemos que se debe al hecho que las visualizaciones de los perfiles con repositorios menores generados mediante *TFIDF* salen las palabras de forma uniforme (Figura 5.5).

5.2.3 Discusión de resultados

Las Figuras 5.3 y 5.4 mostradas forman parte de los resultados del *usuario 25* y con el cual vamos a exponer algunas observaciones del proceso de enriquecimiento. El repositorio utilizado para la visualización del *TFIDF* tiene un tamaño de 2118 palabras generadas de 108 páginas web con las que cuenta el usuario. Este reposito-

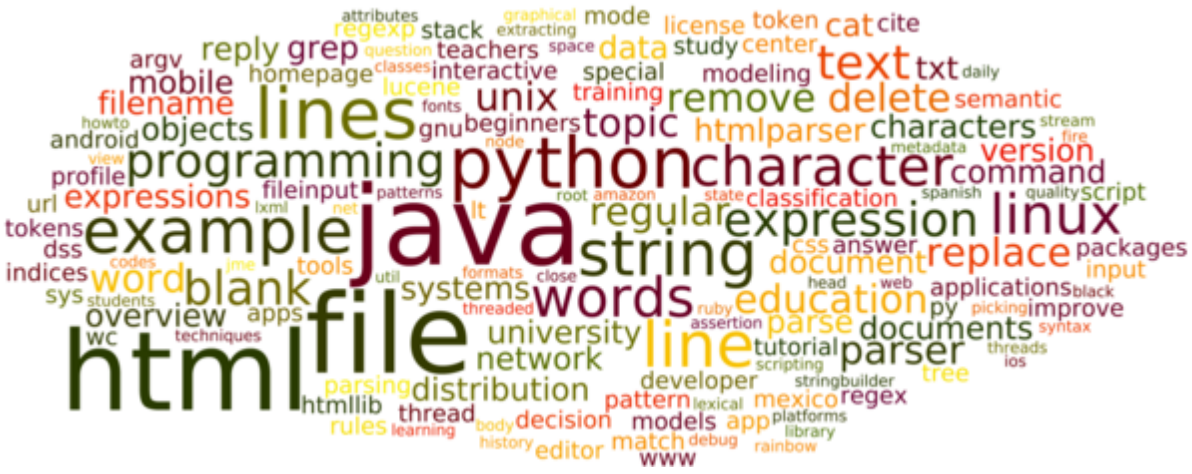


Figura 5.3: Nube de palabras procesadas mediante TFIDF.

Figura 5.4: Nube de palabras procesadas mediante TFIDF y enriquecida con WordNet.

rio *TFIDF* una vez que ha sido pasado por el mecanismo de enriqueciemiento, logra alcanzar un tamaño de 15500 palabras. En la Figura 5.3 podemos observar algunas palabras como, *regular, command, code, line, training* y *license*. Estas palabras al ser utilizado el mecanismo de enriquecimiento, agregan palabras que no corresponden a intereses del usuario (como son, *military, armed* y *war*). De la misma manera son agregadas palabras como, *communication, speech, written, writing, spoken* y *composition*; todas ellas agregadas principalmente por la palabra *language*. Si observamos en la Figura 5.3 el interés por los lenguajes de programación es una de las cosas que más se destaca de la visualización, pero este grupo de palabras las cuales liderea *communication*, no parecen darnos un interés del usuario como las encontradas en la visualización del *TFIDF*. Palabras en la Figura 5.4 como, *biological, chemical, genetic, jurisprudence* y *golf*, son palabras con las cuales el usuario no siente interés alguno.

En las palabras mencionadas anteriormente (*regular, command, code, line, training* y *license*) *WordNet* toma el significado que representan en el contexto militar y es por eso que son agregadas las palabras *military, armed* y *war*, pero al mismo tiempo también son agregadas las palabras *science, computer, computing* y *system*, que representan el contexto de la informática o de las ciencias computacionales (utilizando las palabras encontradas en la visualización, como son: *computer science*). Este fenómeno de múltiples significados es conocido como *polisemia* y es natural que se encuentre en *WordNet* debido a que *WordNet* es una entidad semántica.

Observando ambas Figuras (5.3 y 5.4) podemos darnos cuenta cómo la frase *computer science* (Figura 5.4) engloba la mayoría de las palabras de la Figura 5.3, esto se debe a la selección de la relación utilizada en el mecanismo de enriquecimiento que fue *Hiperónimos*.

Como siguiente ejemplo podemos observar las Figuras 5.5 y 5.6 que aún y que pertenecen al mismo usuario (*Usuario 2*, tamaño de repositorio de 5 documentos) los mecanismos utilizados para generar las visualizaciones nos muestran resultados que parecerían ser de usuarios diferentes. Si observamos ambas figuras podemos notar

Figura 5.5: Nube de palabras generada con TFIDF de repositorio pequeño.

que a la vista parece más enriquecedora la visualización de la Figura 5.6, debido a que los tamaños y la variedad de palabras es mucho mayor que en la Figura 5.5. La votación promedio de la Tabla 5.10 para las visualizaciones por *frecuencia de palabras* puede ser un indicador de la observación que acabamos de señalar.

Figura 5.6: Nube de frecuencia de palabras de repositorio pequeño.

5.3 Recuperación de grupos de usuarios

Sabemos que los usuarios guardan información personal en sus repositorios y hemos hecho una extracción de palabras que reflejan sus intereses. Con esta informa-

ción podemos relacionar a los usuarios, es decir, existe información en sus repositorios que tienen en común y puede ser utilizada para agrupar usuarios por preferencias.

Con los resultados de los perfiles de los usuarios podemos observar (mediante *matrices de similitud*) que existe una relación entre los usuarios, pero no sabemos qué tanta.

5.3.1 Matrices de similitud

Las matrices de similitud son una herramienta utilizada principalmente para inspeccionar agrupamientos de manera visual (Tan *et al.*, 2007). En éstas, los datos se reordenan de tal suerte que aquellos elementos pertenecientes a un mismo grupo quedan contiguos. La similitud entre un par de elementos toma un color o sombreado; por lo general, un tono más oscuro indica un mayor grado de similitud (siendo así que una celda negra indica la similitud máxima de 1.0 y una celda blanca la mínima de 0.0), pero este esquema también puede invertirse. Ya que la similitud entre los elementos de un mismo grupo (*cohesión*) debe ser mayor que la similitud con respecto a otros grupos (*separación*), una *diagonal principal en bloques* debe ser claramente distinguible en la matriz. Cuando esta no es visible, se considera que los grupos tienen poca cohesión. El caso ideal lo ilustra la Figura 5.7.

5.3.2 Configuración

Sabemos que los usuarios son estudiantes de una misma licenciatura (Ingeniero en Tecnologías de Software) y un mismo semestre (quinto), dentro de este mismo grupo se encuentran los usuarios de maestría, quienes cursan un posgrado relacionado también con *ciencias computacionales*; por tanto, se esperará que sus intereses sean similares. Por el contrario, no se espera que los intereses de la muestra restante (los matrimonios jóvenes) con respecto a este grupo sean similares, así como también se espera que la relación fuera mucho más débil.

Con la información de los perfiles podemos, utilizar la *similitud cosenoidal*

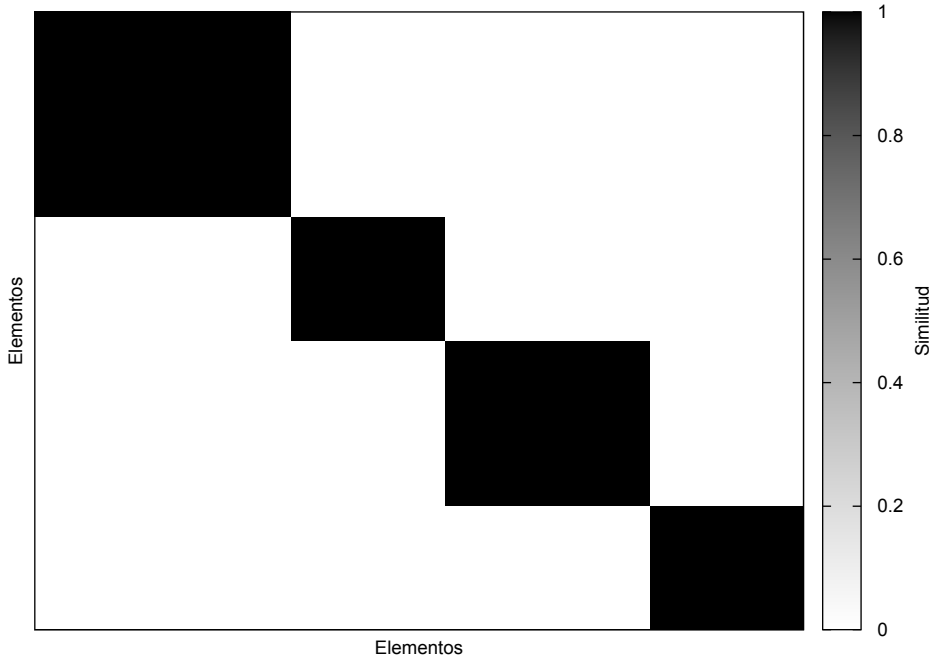


Figura 5.7: Muestra de caso ideal de cohesión.

(Ecuación 2.2) de los repositorios para encontrar la relación que existe entre los usuarios. Con los valores de la *similitud cosenoidal* podemos observar (mediante *matrices de similitud*) los usuarios que se relacionan e incluso podemos agruparlos para hacerles recomendaciones de intereses en común.

Debido a que los perfiles obtenidos cuentan con un extenso vocabulario (especialmente los enriquecidos), se trabajó solamente con las 20 palabras más frecuentes de cada perfil. También, dado que había una considerable diferencia entre las frecuencias máximas de los perfiles (en algunos eran muy altas y en otros muy bajas), se optó por elevar al cubo las frecuencias de aquellos perfiles cuyo máximo era de 6. Para ilustrar lo anterior, si la palabra más frecuente del perfil P_a tenía 4 repeticiones, se agregaban otras 60 al perfil ($60 + 4 = 64 = 4^3$).

Para calcular la similitud entre pares de perfiles (*i.e.* el valor de las celdas de la matriz), cada uno de estos fue representado como un *vector de pesos TFIDF*. Para obtener estos pesos, cada perfil fue tomado como un documento e introducido a un indizador (*Apache Lucene*); la colección de documentos, como podemos intuir,

correspondía a los perfiles de todos los usuarios. La métrica utilizada fue la de *similitud cosenoidal* (2.2).

5.3.3 Resultados

En la Figura 5.8 podemos observar los valores *TFIDF* de la colección completa distribuidos en dos grupos, el grupo principal es el de los estudiantes y podemos ver que aún y con valores fríos, existe relación entre ellos. En la Figura 5.9 observamos los valores de los *perfiles enriquecidos* y vemos cómo el contenido del enriquecimiento enfatiza la tendencia mostrada en la Figura 5.8.

Debido a que los perfiles provenientes de repositorios pequeños (mismos que tenían frecuencias máximas bajas originalmente) arrojaron valores de similitud también muy pequeños con respecto a los elementos de su mismo grupo, se repitió la prueba, pero ahora eliminando aquellos perfiles cuyas frecuencias máximas eran menores a 10. Esto redujo la colección de documentos a 19 perfiles. Las Figuras 5.11 y 5.10 muestran los resultados y se puede apreciar la disminución de ruido. Todos los valores de las figuras de colección completa y reducida pueden ser localizdos en el Apéndice C.

En las Figuras 5.12 y 5.13 observamos los tres grupos de la muestra (izquierda a derecha, licenciatura, posgrado y restante) y podemos ver que los grupos siguen siendo cohesivos. Si eliminamos los usuarios con contenido pequeño, las Figuras 5.14 y 5.15 nos muestran cómo los resultados son mucho mejores.

La Tabla 5.11 muestra los resultados de la *similitud cosenoidal* entre los usuarios de licenciatura con perfiles enriquecidos, tomando el cero como el resultado en el que nada se relaciona entre ellos y el uno como el valor en donde todo se relaciona; este valor sólo lo alcanza el mismo usuario.

Como podemos ver en la Tabla 5.11 los valores de las relaciones entre los usuarios enfatizados con **negrita** son valores que empiezan apartir del .5 y son valores dentro de la *similitud cosenoidal* que son considerados como buenos. Aquellos valores

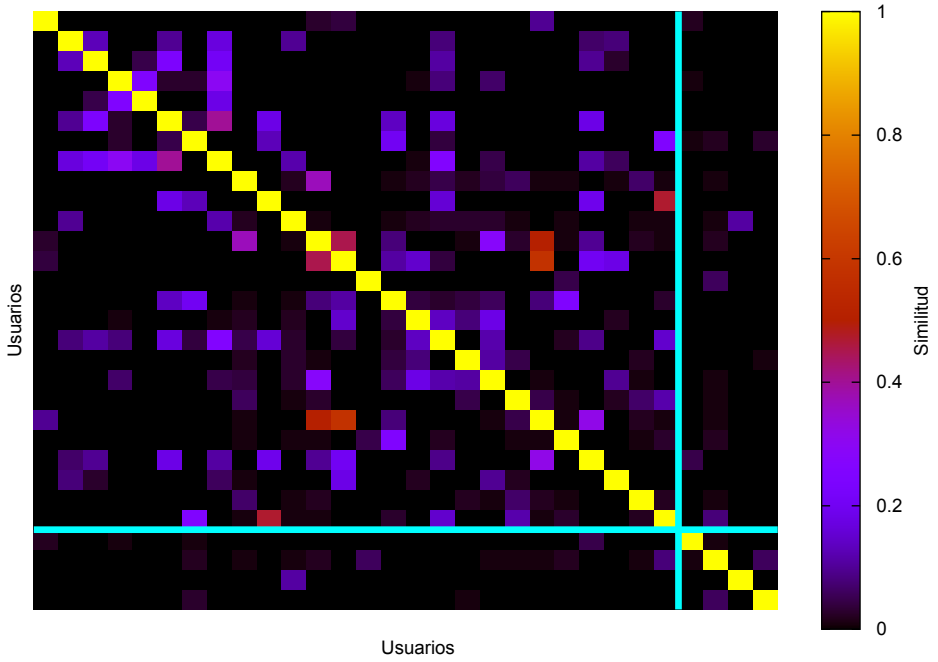


Figura 5.8: TFIDF colección completa, 2 grupos.

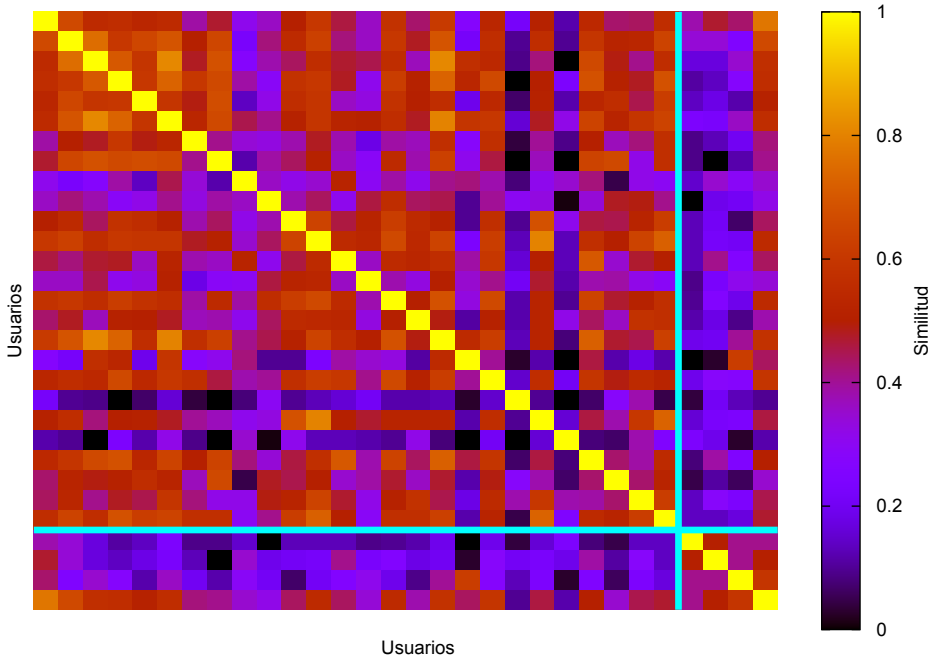


Figura 5.9: Enriquecidos colección completa, 2 grupos.

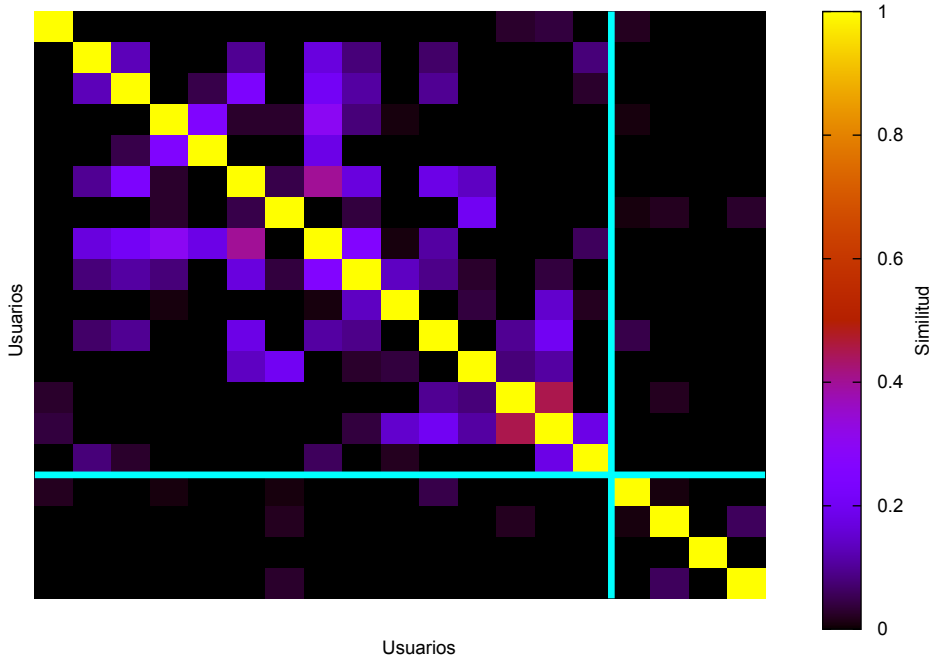


Figura 5.10: TFIDF colección reducida, 2 grupos.

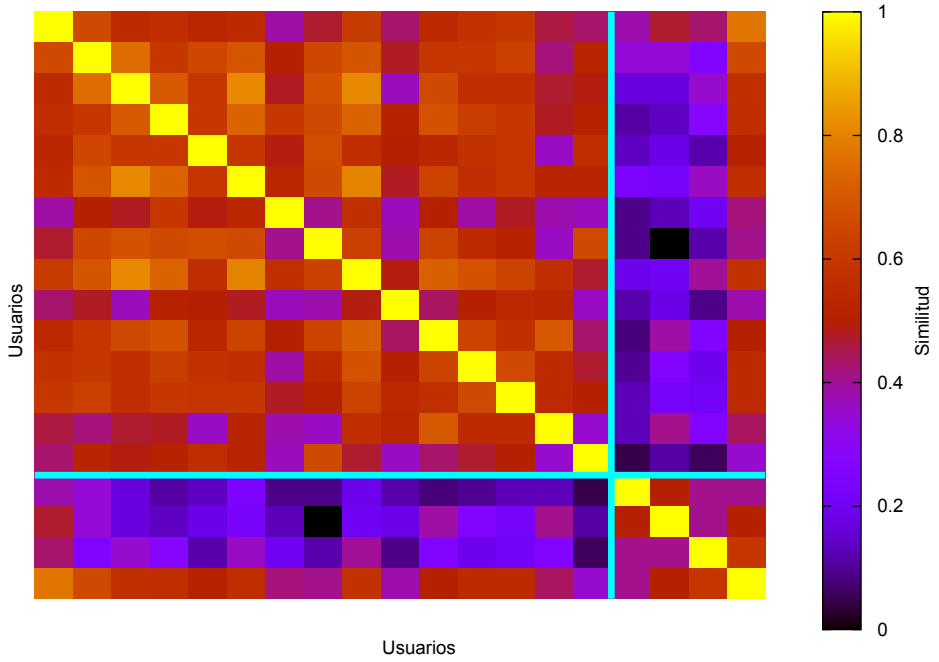


Figura 5.11: Enriquecidos colección reducida, 2 grupos.

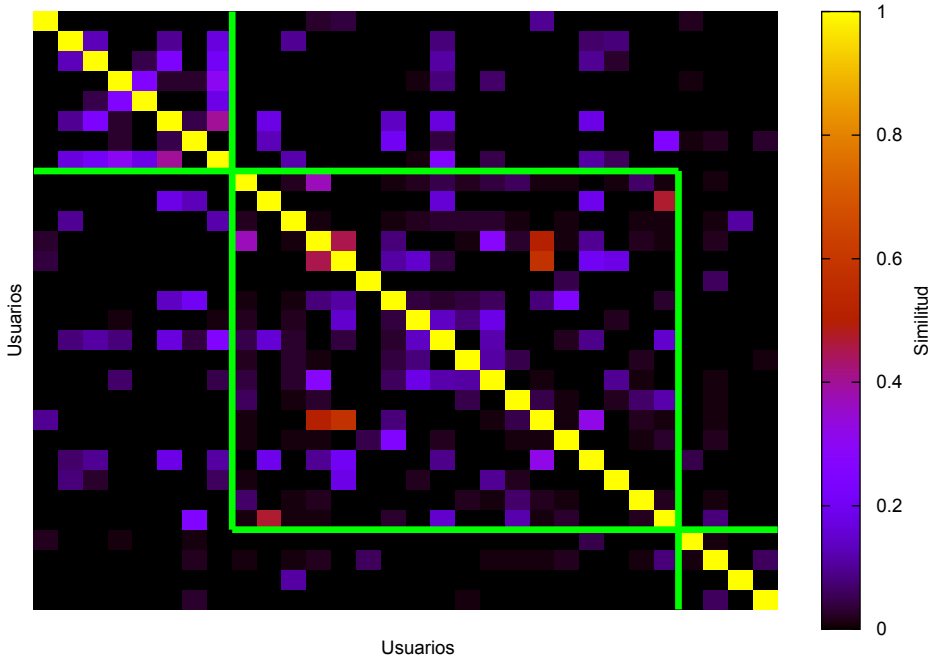


Figura 5.12: TFIDF colección completa, 3 grupos.

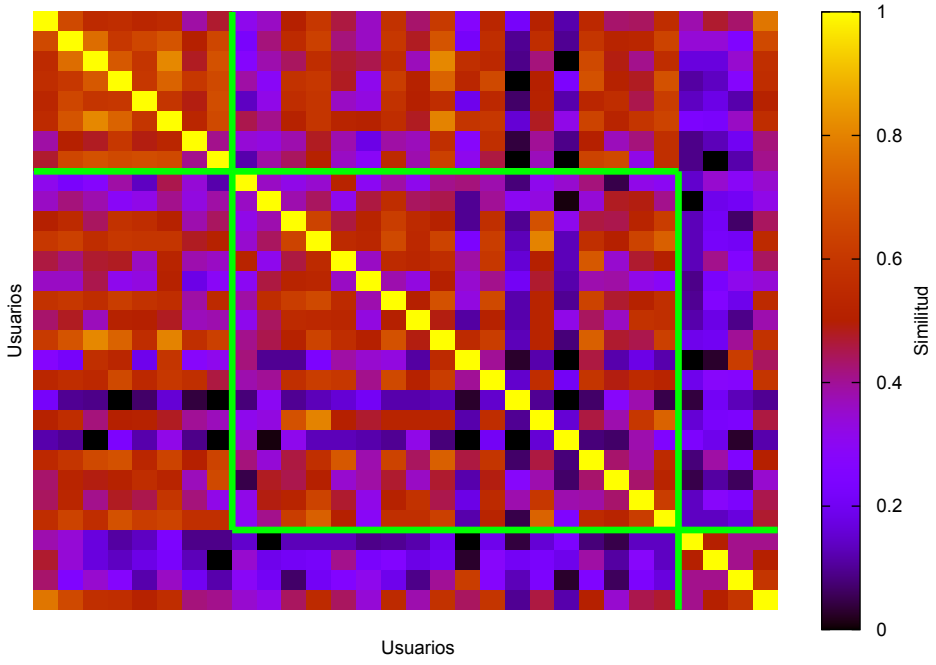


Figura 5.13: Enriquecidos colección completa, 3 grupos.

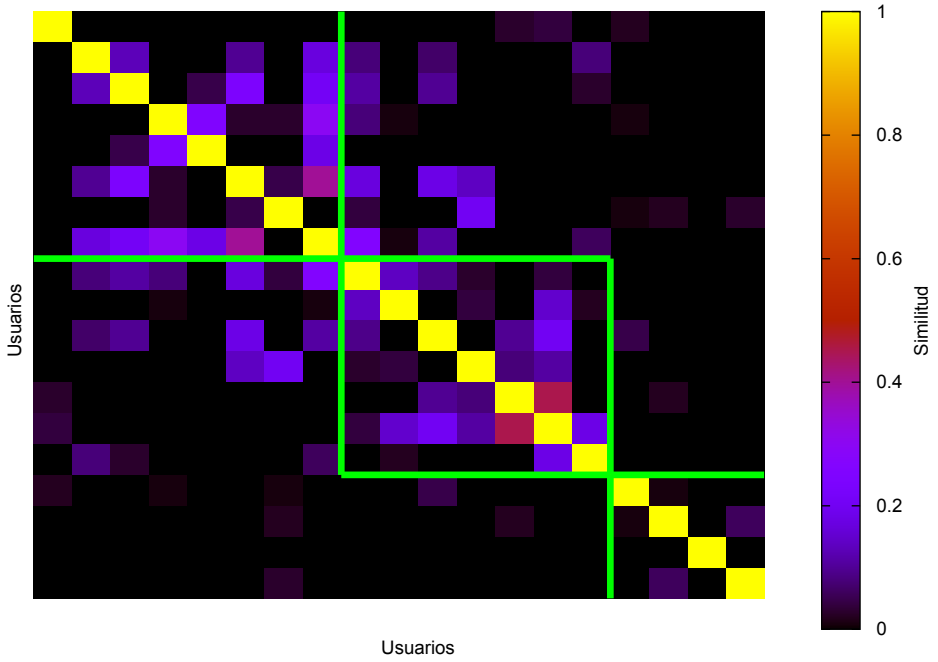


Figura 5.14: TFIDF colección reducida, 3 grupos.

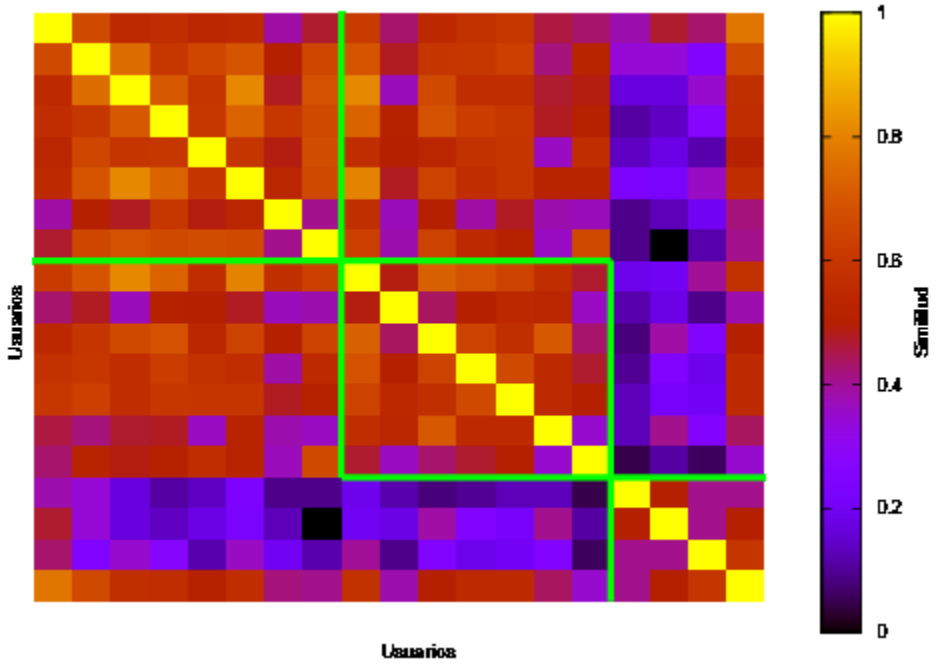


Figura 5.15: Enriquecidos colección reducida, 3 grupos.

	U 16	**U 14**	**U 23**	**U 13**	**U 4**	**U 10**	**U 24**	**U 30**
U 16	**1**	0.49	**0.72**	**0.68**	**0.64**	**0.56**	0.47	**0.63**
U 14	0.49	**1**	0.44	**0.5**	**0.54**	**0.53**	0.36	0.38
U 23	**0.72**	0.44	**1**	**0.64**	**0.57**	**0.7**	0.43	**0.64**
U 13	**0.68**	**0.5**	**0.64**	**1**	**0.66**	**0.55**	0.47	**0.55**
U 4	**0.64**	**0.54**	**0.57**	**0.66**	**1**	**0.55**	**0.5**	**0.51**
U 10	**0.56**	**0.53**	**0.7**	**0.55**	**0.55**	**1**	0.35	0.36
U 24	0.47	0.36	0.43	0.47	**0.5**	0.35	**1**	**0.66**
U 30	**0.63**	0.38	**0.64**	**0.55**	**0.51**	0.36	**0.66**	**1**

Tabla 5.11: Valor de la relación entre los usuarios licenciatura.

enfatizados con *cursiva* son valores desde .49 hacia abajo, que representan intereses del contenido de los repositorios entre dos usuarios que se relacionan; aunque en menos cantidad.

	U 5	**U 12**	**U 15**	**U 6**	**U 18**	**U 25**	**U 26**
U 5	**1**	**0.66**	**0.55**	**0.56**	**0.53**	**0.55**	0.39
U 12	**0.66**	**1**	**0.75**	**0.6**	**0.65**	**0.69**	**0.5**
U 15	**0.55**	**0.75**	**1**	**0.7**	**0.59**	**0.81**	0.48
U 6	**0.56**	**0.6**	**0.7**	**1**	**0.6**	**0.73**	**0.6**
U 18	**0.53**	**0.65**	**0.59**	**0.6**	**1**	**0.59**	0.49
U 25	**0.55**	**0.69**	**0.81**	**0.73**	**0.59**	**1**	**0.53**
U 26	0.39	**0.5**	0.48	**0.6**	0.49	**0.53**	**1**

Tabla 5.12: Valor de la relación entre los usuarios maestría.

En la Tabla 5.12 podemos observar los valores superiores a .5 son más frecuentes en las relaciones de los usuarios de maestría que en los de licenciatura. Todos los usuarios parecen tener buena relación entre el contenido de sus repositorios.

La Figura 5.16 muestra los resultados que relaciona a los usuarios de la muestra investigada, el orden de la muestra en ambos ejes es: usuarios de maestría, usuarios de licenciatura y el grupo restante. Los colores más cálidos respresentan una relación

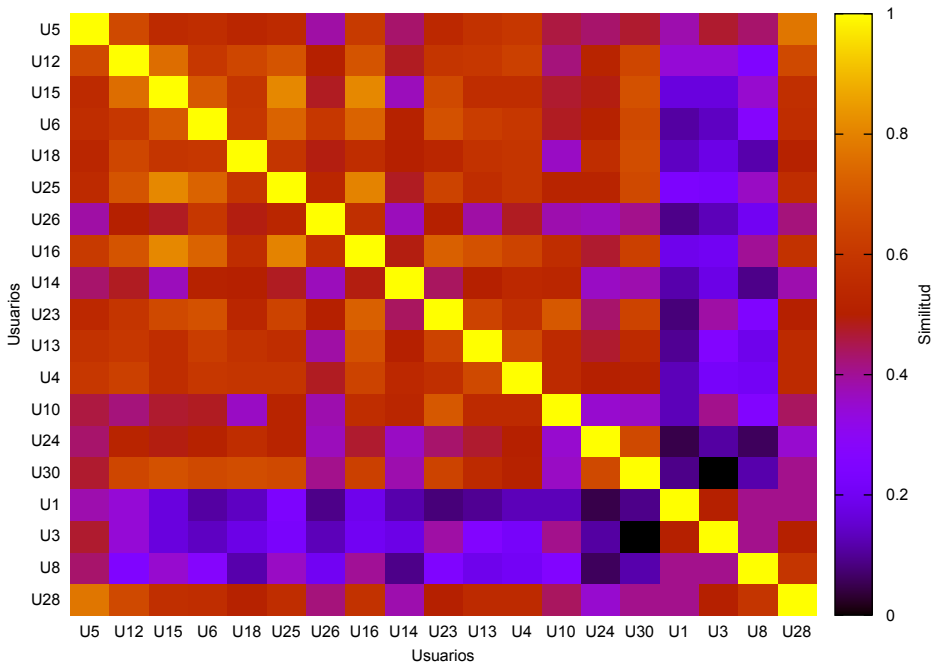


Figura 5.16: Resultados de relación entre usuarios de perfiles enriquecidos.

alta entre sus repositorios y los colores fríos representan poca relación. Los datos utilizados para la generación de esta gráfica fueron los obtenidos a través de la *similitud cosenoidal* que fue calculada tanto para los perfiles con *TFIDF* como para los enriquecidos. Podemos observar que la mayoría de la muestra tiene colores cálidos lo cual nos puede hacer sentido debido a que la mayoría de la muestra se encuentran en el área de computación, tanto en licenciatura como a nivel posgrado. La franja en colores fríos es en donde se encuentran los usuarios de la muestra que no pertenecen a la muestra estudiantil y es importante señalar como el *usuario 8* (U8) que pertenece al grupo restante, tiene relación con el grupo de licenciatura y el grupo de maestría. Los colores cálidos del *usuario 28* (U28), son más comunes con el grupo de maestría en donde sólo el *usuario 25* (U25) tiene color morado. El valor de la relación según podemos observar en el indicador de la derecha de la Figura 5.16 debe de andar rondando el .4. El *usuario 28* tiene sólo tres colores cálidos (U14, U23 y U13) en el grupo de licenciatura pero estos deben andar rondando el valor .5 al igual que los usuarios de su mismo grupo (U3 y U8).

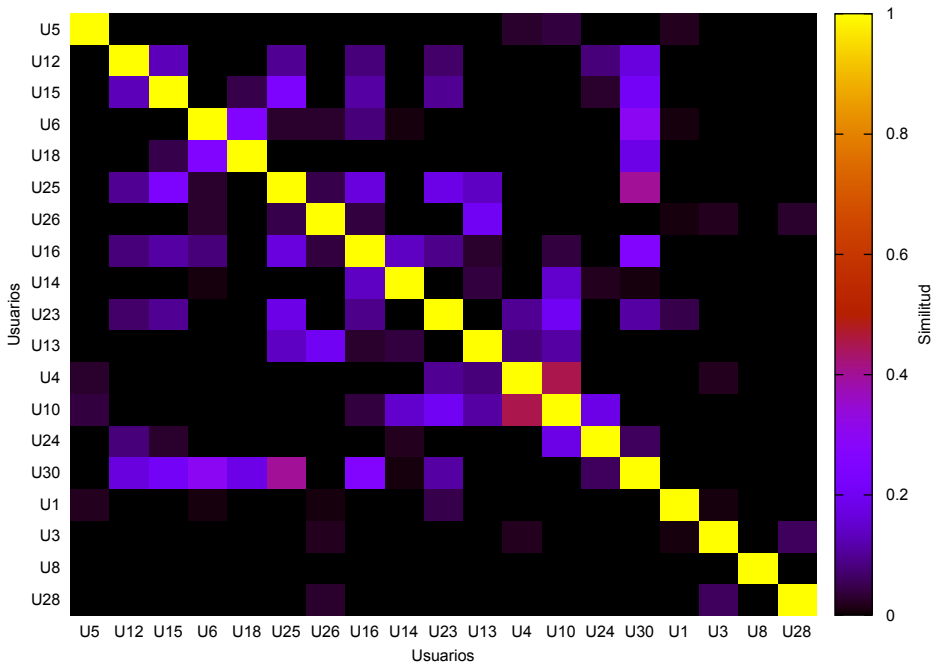


Figura 5.17: Resultados de relación entre usuarios de perfiles generados con TFIDF.

Generamos la Figura 5.17 para ver el comportamiento de las relaciones entre los usuarios con perfiles generados mediante *TFIDF* sin embargo los valores de la *similitud cosenoidal* salen muy cercanos a cero. Si observamos en el indicador de la derecha de ambas figuras (5.16 y 5.17), pareciera que los valores se han corrido para tener como máximo el .5 y no el 1. Este fenómeno lo podemos entender si recordamos cómo funciona el mecanismo de enriquecimiento a través de *WordNet*, lo que hacemos es agregar palabras a los repositorios de los usuarios que han sido encontradas a través de su relación son los *synsets*. Esas mismas palabras son las que enfatizan las relaciones de los usuarios en la Figura 5.16, pero el comportamiento es igual en ambas figuras sin embargo en la Figura 5.17 no lo podemos apreciar.

La cohesión total se calcula como el promedio de las cohesiones grupales, donde la cohesión grupal es el promedio de la similitud entre elementos del mismo grupo. La separación total, análogamente, se calcula como el promedio de las separaciones grupales, donde la separación grupal es el promedio de la similitud entre elementos de grupos distintos.

	Cohesión	**Separación**	**Cohesión/Separación**
TFIDF completo	0.18	0.01	19.61
TFIDF reducido	0.22	0.01	17.62
Enriquecido completo	0.56	0.33	1.69
Enriquecido reducido	0.62	0.37	1.69

Tabla 5.13: Valores promedio de cohesión y separación.

5.3.4 Discusión de resultados

Es importante mencionar que el grupo de maestría tiene en total 419 documentos web procesados y el grupo de licenciatura tiene 217, esto puede ser un indicador que nos muestre que los resultados pueden ser mejores entre más documentos tengamos.

	U 1	**U 3**	**U 8**	**U 28**
U 1	**1**	**0.5**	0.41	0.41
U 3	**0.5**	**1**	0.41	**0.5**
U 8	0.41	0.41	**1**	**0.59**
U 28	0.41	**0.5**	**0.59**	**1**

Tabla 5.14: Valor de la relación entre los usuarios fuera de la muestra estudiantil.

El grupo restante a pesar de ser un muestra menor a la de licenciatura, su promedio en cantidad de documentos por usuario es de 39.25 superior al promedio del grupo de licenciatura que fue de 27.12, es decir, contenían más documentos a procesar. En los resultados de la muestra restante no encontramos valores por encima del .6, aún así los valores mostrados en la Tabla 5.14 no son malos y encontramos relación entre ellos.

A pesar de que los valores de la *similitud cosenoidal* considerados como buenos están establecidos a partir de .5, los valores por debajo pueden ser considerados entre usuarios. Si tenemos un valor de .3 entre dos usuarios, el valor no es malo simplemente nos dice que entre ambos contenidos existe menos información en común, pero aún

el .3 de relación puede ser útil en alguna aplicación.

5.3.5 Discusión

Las gráficas termales (*matrices de similitud*) nos han permitido ver cómo los tamaños en los repositorios afectan el comportamiento y como la utilización de un mecanismo de enriquecimiento enfatiza el comportamiento mostrado por los perfiles con *TFIDF*. Las relaciones que existen entre los estudiantes también resultan claras y aquellos usuarios que no se comportan como la mayoría son aquellos con repositorios pequeños o con intereses en su repositorio de información, diferentes al grupo.

5.4 Discusión general y resumen del capítulo

En la experimentación era importante observar como se comportaban los usuarios con las visualizaciones de sus perfiles. De la muestra utilizada identificamos a los usuarios de licenciatura, los usurios de maestría y la muestra restante. El grupo de los usuarios de licenciatura y maestría representan a un mismo contexto que es el de ser estudiantes en el área de ciencias computacionales de la Facultad de Ingeniería Mecánica y Eléctrica, la muestra restante no forma parte de este contexto. También se identificaron dos grupos, que son aquellos con repositorios grandes y los de repositorios pequeños; esto para observar su comportamiento en la evaluación.

Las evaluaciones realizadas mostraron que los usuarios sintieron que las visualizaciones generadas con *TFIDF* (votación 4.52) contenían más palabras que reflejaban sus intereses, seguido de las visualizaciones de *frecuencia de palabras* (votación 3.79), *perfil enriquecido* (votación 3.38) y *Dummy* (votación 2.28). Las votaciones para las visualizaciones de *frecuencia de palabras* y *perfil enriquecido* están por encima de lo *Regular* (3 estrellas), de tal forma que no consideramos que sean resultados negativos ya que los usuarios sienten interés con las palabras en estas visualizaciones. El *Dummy* salió un poco por encima de *Malo* (2 estrellas), el resultado es el esperado, sin embargo hubo usuarios que calificaron con *Muy Bueno* (5 estrellas)

la visualización. Se les cuestionó a algunos de ellos y mencionaron que la visualización contenía temas que les interesaban mucho, sin embargo ninguno de estos temas fueron encontrados en sus repositorios de información personal.

Nube	**Rep. Pequeño**	**Rep. Grande**
TFIDF	4.416	4.6
Enriquecido	3.33	3
TF	3.916	3.77

Tabla 5.15: Comparativa de valores promedio de los grupo de usuario con menor y mayor número de documentos.

Los grupos de repositorios grandes y repositorios pequeños mantuvieron el orden en las votaciones y sobre todo el grupo de repositorios de mayor tamaño, mantuvo valores muy cercanos a la muestra total como se aprecia en la Tabla 5.15. En el grupo de los repositorios de menor tamaño las visualizaciones con *TFIDF* mostraban las palabras con un tamaño uniforme, pensamos que esto iba a ser un factor en la evaluación, pero el resultado nos muestra que no fué así. Las visualizaciones de *frecuencia de palabras* salieron con un valor muy cercano a *Bueno* (4 estrellas), lo cual nos indica que las palabras que fueron discriminadas durante el proceso de normalización bajo el mecanismo de *TFIDF*, enriquecieron la visualización de los repositorios pequeños. Las visualizaciones de los *perfiles enriquecidos* obtuvieron una votación por encima de *Regular* (3 estrellas), lo cual nos da evidencia que el mecanismo de enriquecimiento propuesto en esta investigación, tuvo un efecto positivo.

También se realizaron experimentos para encontrar qué tanta relación existía entre los perfiles de los usuarios, esto como un análisis de recuperación de usuarios. Los grupos (licenciatura, maestría y restante) fueron utilizados con su información de *perfiles enriquecidos* y perfiles con *TFIDF*, las relaciones de los enriquecidos mostraron relaciones por encima del .5 en su mayoría; lo cual es un resultado bueno en *similitud cosenoidal*. Los valores en las relaciones de los perfiles con *TFIDF* en su mayoría estuvieron cercanos al cero y como valores máximos cercanos al .5, esto

no es malo debido a que nos sirvió como evidencia para ver como el mecanismo de enriquecimiento enfatizó las relaciones de los usuarios.

En el Capítulo 6 mencionamos algunas opciones como trabajos futuros en esta misma línea.

Capítulo 6

Conclusiones y trabajos futuros

6.1 Resumen de lo propuesto en la tesis

Durante esta investigación propusimos generar un perfil de usuario desde un repositorio de información personal para poder recuperar información del usuario que pudiéramos reconocer como sus intereses. Tomando un repositorio de información del usuario y bajo técnicas de *minería de datos* logramos limpiar el repositorio de *palabras vacías* y *listas negras*. Teniendo este repositorio en un *saco de palabras*, generamos los perfiles de los usuarios utilizando un mecanismo de pesos como el *TFIDF* que nos permitió sacar las palabras más relevantes del repositorio. Generamos las visualizaciones de los perfiles con *TFIDF* de los usuarios de la muestra a través de una *nube de palabras* y las visualizaciones de los perfiles con escasa información salían desfavorecidas en comparación de las visualizaciones de perfiles con mucha más información. De tal manera que para hacer frente a este escenario, propusimos utilizar una *entidad semántica* para relacionar sus palabras con las contenidas en los perfiles de los usuarios para enriquecer los perfiles.

Teniendo los perfiles de los usuario con *TFIDF* y los enriquecidos, quisimos saber qué tan relacionados estaban los usuarios a través de sus perfiles; utilizamos la *similitud cosenoidal* para saber el valor. Los valores nos mostraron que existía una buena relación entre los perfiles.

6.2 Respuestas a las preguntas de investigación

Las preguntas realizadas al inicio de esta investigación, fueron todas contestadas durante el proceso de experimentación y son las siguientes:

¿Es posible construir un perfil de usuario a partir de un repositorio de información personal?

Si es posible construir un perfil de usuario a partir de un repositorio de información personal.

¿Cómo representar los intereses del usuario?

La representación de los intereses es representada como un perfil de usuario.

¿Es posible enriquecer el perfil del usuario a partir de su contenido?

Si es posible enriquecer el perfil de usuario utilizando alguna entidad externa bien organizada y con reglas bien establecidas.

¿Cómo visualizar el perfil del usuario?

Nuestro objetivo era que las palabras tuvieran características visuales que representaran su importancia dentro del repositorio personal y la nube de palabras cumple con el objetivo.

6.3 Contribuciones

A continuación enlistamos las contribuciones que se hicieron en esta investigación:

- Modelo conceptual.
- Marco formal de trabajo.
- Listado de palabras a limpiar, generado a partir de las palabras que los mane-

jadores de contenido repiten.

- Procesamiento de repositorios que consisten en marcadores de Internet.
- Proceso para enriquecer el perfil de usuario utilizando una fuente externa de información (*Wordnet*).

6.4 Comentarios conclusivos

Los resultados de la investigación muestran que los usuarios se sintieron identificados con los resultados en las visualizaciones de los perfiles; las visualizaciones del *TFIDF* y de la *frecuencia de las palabras* salieron mejor favorecidas que las visualizaciones de los *perfiles enriquecidos*. Sin embargo, los resultados de los *perfiles enriquecidos* no fueron ajenos del todo y su valor de *Regular* a la percepción de los usuarios, nos deja un resultado no negativo ante el mecanismo propuesto.

Concluimos que los resultados de la investigación nos muestran que debemos seguir trabajando en experimentos para encontrar cifras más absolutas. Es necesario probar más mecanismos que puedan enriquecer el proceso de perfiles y explorar los nuevos escenarios a los que los usuarios se enfrentan hoy en día como los dispositivos móviles y la actividad de redes sociales.

6.5 Trabajos futuros

Como trabajo futuro en el escenario de repositorios pequeños de información (considerando que sean documentos web como en esta investigación) proponemos utilizar un mecanismo que no sólo extraiga las páginas web indicadas por los marcadores, sino también aquellas que forman parte de los enlaces de las primeras. Por ejemplo, si una página p_1 tiene enlaces a otras dos, p_2 y p_3, estas dos también formarían parte del repositorio del usuario. Esto partiendo de la hipótesis de que una

página web contiene material que esta relacionado con el material de los enlaces web con los que cuenta.

En el área de depuración de los documentos web, el mecanismo que se utilizó en esta investigación con respecto a los manejadores de contenido *CMS*; debe de ser ampliado. Las listas utilizadas con las palabras frecuentes que los manejadores de contenido agregan, fueron desarrolladas con sólo el contenido de los repositorios que se utilizaron en esta investigación. Pero debido al uso cada vez más frecuente en Internet de estos sistemas, consideramos que su uso debe de ser considerado para investigación y medir su impacto.

En el área de la semántica es claro que las entidades que más aportan al conocimiento son las realizadas en el idioma inglés, pero es necesario aportar conocimiento a nuestro idioma. La lengua no debe de ser una barrera al conocimiento y los trabajos en español deben de incrementarse. Como trabajo futuro creemos que los esfuerzos a crear entidades como *WordNet* o ayudar a mejorar las que se encuentren en desarrollo, deben de ser considerados.

Apéndice A

Palabras vacías (StopWords)

a able aboard about above abovementioned abovesaid abstract accordance according across actual additional afoot aforementioned aforesaid aforethought aforetime after afterward afterwards again against ago ahead albeit alike all almost alone along alongside already also although altogether always am amid amidst among amongst an and another anti any anybody anyhow anymore anyone anyplace anything anytime anyway anyways anywhere apart apiece apparatus application apr april are around art as aside assignee at atop aug august away b back background be became because become been before beforehand beforementioned beforeseen beforetime began begin begun behalf behest behind being below beneath beside besides best betcha better between beyond big billion billionth both brief bring brought but by c came can cannot capability cc cdt certain cetera characteristic claim classification cm combination come common component comprise comprising concerning configuration consist conventional corresponding could coulda couldn couldnt cross cstd day dec december define dependent describe described description design desired despite detail detailed development device did didn different disclose disclosure do does doesn doesnt doing done dont down dr drawing drawings during e each early edt effective eg eight eighteen eighteenth eighth eightieth eighty either element eleven eleventh else elsewhere embodiment embodiments enough entire entirety equipment equivalent est et etc even ever every everybody everyone everything everywhere ex except existing extend extent f feb february few ff field fifteen fifteenth fifth fiftieth fifty fig figs find first five for foregoing forever form former forthcoming forthright fortieth forty four fourteen

fourteenth fourth fri friday from function further furthermore g get given go gone gonna got gotta gotten h had hadn hadnt hafta half has hasn hasnt have havent having he hence henceforth her here hereabout hereabouts hereabove hereafter hereagain hereas hereat herebefore herebelow hereby herefor herefore herefrom herein hereinafore hereinafter hereinbefore hereindescribed hereinlater hereinto hereinunder hereof hereon hereto heretobefore heretofor heretofore heretoforeknown hereunder hereunto hereupon herewith herewithal herewithin hers herself him himself his hither hitherto hour how howbeit however hrs hundred hundredth i ie if ii iii important improvement in inasmuch inc include including indeed inside insofar instead inter into invention inventor inward inwards irrespective is isnt issue it its itself ive ix j jan january jul july jun june just k kg kind kinda km known l last later latter lb lbs least less lest let li lii liii like likewise limitation liv lix ltd lvi lvii lviii lxi lxii lxiii lxiv lxix lxv lxvi lxvii lxviii lxx lxxi lxxii lxxiii lxxiv lxxix lxxv lxxvi lxxvii lxxviii lxxx lxxxi lxxxii lxxxiii lxxxiv lxxxix lxxxv lxxxvi lxxxvii lxxxviii m make manner many mar material may maybe mdt me means meantime meanwhile mechanism mere method mg mi might mighta mightn mightnt million millionth mins minute ml mm modification mon monday month more moreover most mr mrs mst much must musta mustn mustnt my myself n nb nd nearby necessary need needn neednt needta neither never nevertheless new nine nineteen nineteenth ninetieth ninety ninth nobody non none nonetheless noone nor not nothing notwithstanding nov november now nowadays nowhere o objective oct october of off oft often oftentimes on once one ongoing only onto operation or other others otherwise ought oughta oughtn oughtnt our ours ourselves out outside outward outwards over own oz p part pat patent pdt per percent perhaps pertain plural plurality pm portion pp pre preceding preferred present previous primary principle prior problem procedure process production provide provision pst publication purpose q quadrillion quadrillionth quarter que quintillion quintillionth quite r rather re reference regarding regardless relate related relevant require requirement respective result s said same saturday say scope sec second secs see seem seeming seldom sep sept september ser seven seventeen seventeenth seventh seventieth seventy several shall she should shoulda shouldn shouldnt since six sixteen sixteenth

sixth sixtieth sixty so solution some somebody someday somehow someone someplace something sometime sometimes somewhat somewhere soon sorta specific specification standpoint still such summary sunday system t take technique ten tenth th than that thats the their theirs them themselves then thence there thereabout thereabouts thereabove thereacross thereadjacent thereafter thereagainst therealong thereamong therearound thereas thereat thereatop therebefore therebehind therebeing therebelow therebeneath therebetween therebeyond thereby therefor therefore therefrom therein thereinafter thereinbetween thereinside thereinto thereof thereon thereonto thereout thereover therepast therethrough therethroughout thereto theretofore theretoward thereunder thereunto thereupon therewith therewithal therewithin these they theyve thing third thirteen thirteenth thirtieth thirty this thither thorough those though thousand thousandth three through throughout thru thu thur thurs thursday thus til to today together told tomorrow too top toward towards trillion trillionth tue tues tuesday twelfth twelve twentieth twenty twice two u unable under underneath unless unlike until unto up upon upper us use useta utc v variation various very vi via vii viii viz vs w wanna want was wasn wasnt we wed wednesday week well went were weve what whatever whatsoever when whence whenever where whereabout whereabouts whereafter whereas whereat wherebetween whereby whereever wherefor wherefore wherefrom wherein whereinafter whereinto whereof whereon wheresoever wherethrough whereto whereunder whereunto whereupon wherever wherewith wherewithal whether which whichever while whilst whither who whoever whom whomever whose why will with within without wk worthwhile would woulda wouldn wouldnt x xc xci xcii xciii xciv xcix xcv xcvi xcvii xcviii xi xii xiii xiv xix xl xli xlii xliii xliv xlix xlv xlvi xlvii xlviii xv xvi xvii xviii xx xxi xxii xxiii xxiv xxix xxv xxvi xxvii xxviii xxx xxxi xxxii xxxiii xxxiv xxxix xxxv xxxvi xxxvii xxxviii y yd yds year yesterday yet you your yours yourself yourselves youve yr z

Apéndice B

Evaluaciones de los usuarios

El orden de las votaciones (de izquierda a derecha) es: visualización con *TFIDF*, visualización *perfil enriquecido*, visualización *frecuencia de palabras* y visualización *Dummy*.

u1	4	2	5	3	**u11**	5	3	3	1	**u21**	5	1	5	1
u2	5	4	4	3	**u12**	5	4	5	4	**u22**	5	4	4	3
u3	5	1	3	1	**u13**	5	3	4	1	**u23**	5	5	3	3
u4	4	3	2	2	**u14**	5	3	3	3	**u24**	2	3	4	1
u5	5	5	4	3	**u15**	3	2	4	2	**u25**	5	3	4	1
u6	3	4	2	1	**u16**	5	2	5	2	**u26**	4	2	2	1
u7	5	5	4	4	**u17**	5	3	3	2	**u27**	5	2	3	1
u8	4	3	2	1	**u18**	4	5	4	3	**u28**	5	4	3	2
u9	4	2	5	1	**u19**	5	3	5	5	**u29**	2	5	4	3
u10	5	2	5	4	**u20**	3	5	5	2	**u30**	4	5	1	2

Tabla B.1: Resultados de las evaluaciones de las visualizaciones de los perfiles de los usuarios.

Apéndice C

Cohesión y Separación

	FIME	**Restante**
FIME	*0.073*	*0.005*
Restante	0.005	*0.26*

Tabla C.1: TFIDF colección completa, 2 grupos

	FIME	**Restante**
FIME	*0.46*	*0.26*
Restante	0.26	*0.6*

Tabla C.2: Enriquecidos colección completa, 2 grupos

	Licenciatura	**Posgrado**	**Restante**
Licenciatura	*0.1*	*0.02*	*0.01*
Posgrado	0.02	*0.19*	*0.003*
Restante	0.01	0.003	*0.26*

Tabla C.3: TFIDF colección completa, 3 grupos

	Licenciatura	**Posgrado**	**Restante**
Licenciatura	*0.43*	*0.44*	*0.25*
Posgrado	0.44	*0.65*	*0.31*
Restante	0.25	0.31	*0.6*

Tabla C.4: Enriquecido colección completa, 3 grupos

	FIME	**Restante**
FIME	*0.116*	*0.003*
Restante	0.003	*0.26*

Tabla C.5: TFIDF colección reducida, 2 grupos

	FIME	**Restante**
FIME	*0.585*	*0.286*
Restante	0.286	*0.6*

Tabla C.6: Enriquecidos colección reducida, 2 grupos

	Licenciatura	**Posgrado**	**Restante**
Licenciatura	*0.209*	*0.032*	*0.002*
Posgrado	0.032	*0.192*	*0.003*
Restante	0.003	0.003	*0.26*

Tabla C.7: TFIDF colección reducida, 3 grupos

	Licenciatura	**Posgrado**	**Restante**
Licenciatura	*0.606*	*0.54*	*0.25*
Posgrado	0..54	*0.648*	*0.310*
Restante	0.264	0.305	*0.6*

Tabla C.8: Enriquecido colección reducida, 3 grupos

Bibliografía

Balabanović, M. y Y. Shoham (1997), «Fab: content-based, collaborative recommendation», *Communications of the ACM*, **40**(3), págs. 66–72.

Basu, C., H. Hirsh, W. Cohen *et al.* (1998), «Recommendation as classification: Using social and content-based information in recommendation», en *Proceedings of the national conference on artificial intelligence*, John Wiley & Sons LTD, págs. 714–720.

Billsus, D. y M. J. Pazzani (1999), «A Hybrid User Model for News Story Classification», .

Billsus, D. y M. J. Pazzani (2000), «User modeling for adaptive news access», *User modeling and user-adapted interaction*, **10**(2), págs. 147–180.

Billsus, D., M. J. Pazzani y J. Chen (2000), «A learning agent for wireless news access», en *Proceedings of the 5th international conference on Intelligent user interfaces*, ACM, págs. 33–36.

Bogárdi-Mészöly, Á., A. Rövid y H. Ishikawa (2013), «Topic Recommendation from Tag Clouds», *Bulletin of Networking, Computing, Systems, and Software*, **2**(1), págs. pp–25.

Bouras, C. y V. Tsogkas (2012), «User Personalization via W-kmeans», .

Breese, J. S., D. Heckerman y C. Kadie (1998), «Empirical analysis of predictive algorithms for collaborative filtering», en *Proceedings of the Fourteenth*

conference on Uncertainty in artificial intelligence, Morgan Kaufmann Publishers Inc., págs. 43–52.

BURKE, R. (2002), «Hybrid recommender systems: Survey and experiments», *User modeling and user-adapted interaction*, **12**(4), págs. 331–370.

CLAYPOOL, M., A. GOKHALE, T. MIRANDA, P. MURNIKOV, D. NETES y M. SARTIN (1999), «Combining content-based and collaborative filters in an online newspaper», en *Proceedings of ACM SIGIR Workshop on Recommender Systems*, tomo 60, Citeseer.

CORNELLA, A. (2000), «¿Cómo sobrevivir a la infoxicación?», en *Trascripción de la conferencia del acto de entrega de títulos de los programas de Formación de Posgrado del año académico 1999-2000*.

DELGADO, J. y N. ISHII (1999), «Memory-Based Weighted Majority Prediction», en *ACM SIGIR'99 workshop on recommender systems*, Citeseer.

DOMINICH, S. (2000), «A unified mathematical definition of classical information retrieval», *Journal of the American Society for Information Science*, **51**(7), págs. 614–624.

DUMAIS, S., E. CUTRELL, J. J. CADIZ, G. JANCKE, R. SARIN y D. C. ROBBINS (2003), «Stuff I've seen: a system for personal information retrieval and re-use», en *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, ACM, págs. 72–79.

FELDMAN, R. y J. SANGER (2006), *The text mining handbook: advanced approaches in analyzing unstructured data*, Cambridge University Press.

GANTZ, J. y D. REINSEL (2010), «The Digital Universe Decade: Are You Ready?», .

GAUCH, S., M. SPERETTA, A. CHANDRAMOULI y A. MICARELLI (2007), *User Profiles for Personalized Information Access*, *Lecture Notes in Computer Science*, tomo 4321, capítulo 2, Springer.

Gentilia, G., A. Micarellib y F. Sciarronea (2003), *InfoWeb: An adaptive information filtering system for the cultural heritage domain*, tomo 17, TAYLOR & FRANCIS INC, págs. 715–744.

Gils, B. V. y E. D. Schabell (2003), «User-profiles for information retrieval», en *Proceedings of the 15th Belgian-Dutch Conference on Artificial Intelligence (BNAIC 03)*.

Hotho, A., A. Nürnberger y G. Paass (2005), «A brief survey of text mining», en *LDV Forum-GLDV Journal for Computational Linguistics and Language Technology*, tomo 20, sn, págs. 19–62.

Jabeen, S., X. Gao y P. Andreae (2012), «Using Wikipedia as an External Knowledge Source for Supporting Contextual Disambiguation», .

Jung, J. J. (2005), «Collaborative web browsing based on semantic extraction of user interests with bookmarks», *Journal of Universal Computer Science*, **11**(2), págs. 213–228.

Kanawati, R. y M. Malek (2002), «A multi-agent system for collaborative bookmarking», en *Proceedings of the first international joint conference on Autonomous agents and multiagent systems: part 3*, ACM, págs. 1137–1138.

Konstan, J. A., B. N. Miller, D. Maltz, J. L. Herlocker, L. R. Gordon y J. Riedl (1997), «GroupLens: applying collaborative filtering to Usenet news», *Communications of the ACM*, **40**(3), págs. 77–87.

Kosala, R. y H. Blockeel (2000), «Web mining research: A survey», *ACM Sigkdd Explorations Newsletter*, **2**(1), págs. 1–15.

Kuflik, T. y P. Shoval (2000), «Generation of user profiles for information filtering», .

Lieberman, H. *et al.* (1995), «Letizia: An agent that assists web browsing», en *International Joint Conference on Artificial Intelligence*, tomo 14, LAWRENCE ERLBAUM ASSOCIATES LTD, págs. 924–929.

Manning, C. D., P. Raghavan y H. Schütze (2008), *Introduction to information retrieval*, tomo 1, Cambridge University Press Cambridge.

Manning, C. D. y H. Schütze (1999), *Foundations of statistical natural language processing*, MIT press.

Marlin, B. (2003), «Modeling user rating profiles for collaborative filtering», en *Proc. 17th Ann. Conf. Neural Information Processing Systems (NIPS?03)*.

Mladenic, D. (1996), «Personal WebWatcher: design and implementation», .

Mobasher, B. (2007), *Data Mining for Web Personalization*, capítulo 3, Sringer.

Mobasher, B., H. Dai, T. Luo y M. Nakagawa (2002), «Discovery and evaluation of aggregate usage profiles for web personalization.», .

Mobasher, B., X. Jin y Y. Zhou (2004), «Semantically enhanced collaborative filtering on the web», *Web Mining: From Web to Semantic Web*, págs. 57–76.

Montebello, M., W. Gray y S. Hurley (1998), «A personal evolvable advisor for WWW knowledge-based systems», en *Proc. of Workshop on Reuse of Web Information at WWW*, tomo 98, Citeseer, págs. 59–69.

Moukas, A. (1996), «Amalthaea: Information Discovery And Filtering Using A Multiagent Evolving Ecosystem», .

Pang, B. y L. Lee (2005), «Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales», en *ANNUAL MEETING-ASSOCIATION FOR COMPUTATIONAL LINGUISTICS*, tomo 43, pág. 115.

Pazzani, M. y D. Billsus (1997), «Learning and revising user profiles: The identification of interesting web sites», *Machine learning*, **27**(3), págs. 313–331.

Pazzani, M. J. (1999), «A framework for collaborative, content-based and demographic filtering», *Artificial Intelligence Review*, **13**(5), págs. 393–408.

PAZZANI, M. J., J. MURAMATSU, D. BILLSUS *et al.* (1996), «Syskill & Webert: Identifying interesting web sites», en *Proceedings of the national conference on artificial intelligence*, págs. 54–61.

PORTER, M. F. *et al.* (1980), «An algorithm for suffix stripping», .

PRETSCHNER, A. (1998), *Ontology Based Personalized Search*, Tesis de Maestría.

P.R.KAUSHIK y D. K. N. MURTHY (1999), «Personal Search Assistant: A Configurable Personal Meta Search Engine», URL `http://ausweb.scu.edu.au/aw99/papers/murthy/paper.html`.

RAMANATHAN, K., J. GIRAUDI y A. GUPTA (2008), «Creating hierarchical user profiles using Wikipedia», *HP Labs*.

RICCI, F., L. ROKACH y B. SHAPIRA (2011), «Introduction to recommender systems handbook», *Recommender Systems Handbook*.

RODRIGUEZ, M., J. M. G. HIDALGO y B. D. AGUDO (2000), «Using WordNet to complement training information in text categorization», en *Proceedings of 2nd International Conference on Recent Advances in Natural Language Processing II: Selected Papers from RANLP*, tomo 97, págs. 353–364.

SCHEIN, A. I., A. POPESCUL, L. H. UNGAR y D. M. PENNOCK (2002), «Methods and metrics for cold-start recommendations», en *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, ACM, págs. 253–260.

SCOTT, S., S. MATWIN *et al.* (1998), «Text classification using WordNet hypernyms», en *Use of WordNet in Natural Language Processing Systems: Proceedings of the Conference*, págs. 38–44.

SEMERARO, G., M. DEGEMMIS, P. LOPS y I. PALMISANO (2005), «WordNet-based user profiles for semantic personalization», en *Proceedings of Workshop on New Technologies for Personalized Information Access (PIA 2005)*, págs. 74–83.

Sieg, A., B. Mobasher y R. Burke (2004), «Inferring users information context from user profiles and concept hierarchies», *Classification, Clustering, and Data Mining Applications*, págs. 563–573.

Sinclair, J. y M. Cardew-Hall (2008), «The folksonomy tag cloud: when is it useful?», *Journal of Information Science*, **34**(1), págs. 15–29.

Smyth, B. y P. Cotter (2000), «A personalised TV listings service for the digital TV age», *Knowl.-Based Syst.*, **13**(2-3), págs. 53–59.

Speretta, M. (2000), *Personalizing Search Based on User Search Histories*, Tesis de Maestría.

Tan, P.-N. *et al.* (2007), *Introduction to data mining*, Pearson Education India.

Teevan, J., S. T. Dumais y E. Horvitz (2005), «Personalizing search via automated analysis of interests and activities», en *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, ACM, págs. 449–456.

Trajkova, J. y S. Gauch (2004), «Improving Ontology-Based User Profiles», .