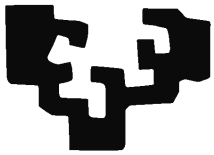

Contribuciones al modelado semántico de las comunidades de práctica en línea

Por

Felipe Aguilera Valenzuela

Depositado en el Departamento de Ciencias de la Computación e Inteligencia Artificial de la Universidad del País Vasco para optar al grado de
Doctor en Informática

Bajo la dirección de:

Prof. Dr. Manuel Graña Romay
Dr. Sebastián Ríos Pérez

Universidad del País Vasco
Euskal Herriko Unibertsitatea
Donostia - San Sebastián

2018

Contribuciones al modelado semántico de las comunidades de práctica en línea

por
Felipe Aguilera Valenzuela

Depositado en el Departamento de Ciencias de la Computación e Inteligencia Artificial de la Universidad del Pais Vasco para optar al grado de Doctor en Informática

Resumen

Esta tesis está dirigida al estudio de un tipo especifico de redes sociales, las comunidades de práctica, que se caracterizan por poseer un interés común que aglutina a los miembros. Consideramos la construcción de los grafos de relación de estas redes sociales en base a las comunicaciones que se realizan entre sus miembros mediante la publicación en los foros internos. Para mejorar la representación eliminando información espuria utilizamos herramientas de modelado semántico. La identificación de los miembros centrales en esta representación mejorada se ajusta mucho mejor a la realidad de la comunidad, usando como información de validación la identificación de miembros relevantes realizada por el administrador de la red. Utilizamos dos comunidades de práctica reales, de las que disponemos tanto de los datos como de la identificación de los miembros relevantes dada por los administradores de la red.

Keywords: *Redes sociales en linea, Comunidad de Práctica, Análisis de redes sociales, Análisis semántico, Latent Dirichlet Analysis.*

.

Agradecimientos

El trabajo de esta tesis ha sido parcialmente soportado por fondos FEDER para el proyecto MINECO TIN2017-85827-P, el proyecto KK-2018/00071 de la convocatoria Elkartek 2018 del Gobierno Vasco, y el proyecto H2020-MSCA-RISE CybSPEED de numero 777720.

Felipe Aguilera Valenzuela

‘

Índice general

1. Introducción **1**

1.1. El contexto general de la revolución digital en las redes sociales 1
1.2. Motivación 5
1.3. Hipótesis y Objetivos de la Tesis 6
1.4. Metodología Utilizada 7
1.5. Casos reales utilizados para la experimentación 8
1.6. Contribuciones de la Tesis 9
1.7. Publicaciones 10
1.8. Estructura de la Tesis 12

2. Estado del arte **13**

2.1. Estructuras Sociales 13
2.1.1. Redes Sociales 14
2.1.2. Comunidades Virtuales 18
2.1.3. Comunidades Virtuales de Práctica 21
2.2. Estrategias de Evaluación de Comunidades Virtuales 32
2.2.1. El Punto de Vista Sicológico 33
2.2.2. El Punto de Vista Sociológico 34
2.3. Técnicas de Evaluación en Redes Sociales y Comunidades Virtuales 35
2.3.1. Análisis de Redes Sociales 36
2.3.2. Análisis de Comunidades Virtuales 39

2.4. Modelado de Aspectos Sociales 41

3. Casos de estudio **47**
3.1. Comunidad Virtual I 47
3.1.1. Análisis de la Comunidad 48
3.2. Comunidad Virtual II 51
3.2.1. Análisis de la Comunidad 51

4. Aproximaciones al modelado estructural/semántico **54**
4.1. Descripción General 55
4.2. Procesamiento del texto 58
4.2.1. Lógica Difusa para Clasificación basada en Conceptos . 59
4.2.2. Proceso de Extracción de Tópicos usando modelos gráficos 61
4.3. Generación de la Red Social 64
4.4. Filtrado de la Red Social basado en Conceptos & Tópicos . . . 67
4.5. Algoritmo de Centralidad 68
4.6. Construcción del grafo pesado de Red Social guiado por información semántica 71

5. Experimentos, Resultados y Evaluación **74**
5.1. Métricas y Metodología de Evaluación 74
5.2. Comunidad I 78
5.2.1. Detección de tópicos y Definición de Conceptos 78
5.2.2. Filtrado de arcos basado en conceptos y tópicos 81
5.2.3. Detección de miembros claves 84

6. Conclusiones y Trabajo Futuro **89**
6.1. Conclusiones 89
6.2. Trabajo futuro 90

Índice de figuras

2.1. Representación gráfica de una red social en forma de un sociograma 15
2.2. Tipos de comunidades virtuales (Figura obtenida de Henri2003). 21
2.3. Evolución de una comunidad de práctica (Fig. obtenida de [Wenger2002]). 24
2.4. Niveles de participación de los miembros de una comunidad (Figura obtenida de [Wenger2002]). 25
2.5. Interrelación de aspectos sociales que posee los sistemas de apoyo a comunidades virtuales. 45

3.1. Actividad de la Comunidad I. 49
3.2. Total de miembros activos de la Comunidad I, basados en el total de interacciones. 51
3.3. Actividad de la Comunidad II. 52
3.4. Total de miembros activos de la Comunidad II, basados en el total de interacciones. 53

4.1. Foco Modelo 56
4.2. Networks Types 64
4.3. Forum Structure 65
4.4. Tres diferentes modelos de representación en forma de red las interacciones dentro de un foro. (a) creator, (b) reply_prev, (c) reply_all 66

5.1. Marco metodológico de evaluación. 75
5.2. Comparación de tiempos de procesamiento entre LDA y PYTM, para el año 2013, con cantidad de tópicos entre 5 y 100, medido en horas. 79
5.3. Comparación de técnicas LDA y PYTM en la Comunidad I, Año 2013, entre 5 y 100 tópicos. 80
5.4. Total de tópicos para cada año de estudio para las técnicas PYTM y LDA. 81
5.5. Historic Network density 82
5.6. NetReduction2013 83
5.7. Comparación de las 4 redes Descubrimiento de miembros claves aplicados a los top 10, 20, 30, 40 miembros ordenaos por HITS hub. 84
5.8. Descubrimiento de miembros claves aplicados a los top 10, 20, 30, 40 miembros ordenaos por HITS hub. 85
5.9. Métodos para descubrir miembros claves aplicados a los top 10, 20, 30, 40 miembros ordenados por HITS hub. 87

Índice de cuadros

2.1. Resumen de los aspectos sociales de las estructuras sociales en estudio, desde la perspectiva de diversos autores. 26

3.1. Participación de los miembros en los foros de discusión de la Comunidad I 49
3.2. Estadísticas de los hilos de discusión de la Comunidad I 50
3.3. Estadísticas de los hilos de discusión de la Comunidad II . . . 53

5.1. Diez de las palabras más relevantes con sus respectivas probabilidades condicionales para dos tópicos obtenidos con LDA. . 80
5.2. Detección de miembros claves para diferentes métodos y configuración de parámetros, para los últimos 5 meses del año 2013, utilizando la red *creator* 86
5.3. Rendimiento de la detección de miembros clave de Tipo A utilizando el método tradicional y la combinación con filtro LDA 88
5.4. Rendimiento de la detección de miembros clave de Tipo $A+B$ utilizando el método tradicional y la combinación con filtro LDA 88
5.5. Rendimiento de la detección de miembros clave de Tipo $A+B+C$ utilizando el método tradicional y la combinación con filtro LDA 88

Capítulo 1

Introducción

En este capítulo damos primeramente una motivación de la tesis repasando el contexto general. A continuación, formulamos la hipótesis de trabajo fundamental que vertebra la tesis doctoral y los objetivos que persigue. Después de introducir la metodología empleada presentamos los casos de estudio sobre los que hemos trabajado en las demostraciones prácticas. Finalmente, presentamos las contribuciones identificadas de la tesis, los resultados en forma de publicaciones que respaldan la tesis y la estructura de la tesis.

1.1. El contexto general de la revolución digital en las redes sociales

Einstein señalaba que en el siglo 20 han sido tres las bombas que han explotado con repercusiones mundiales: la demográfica, la atómica y la de telecomunicaciones [Levy2001]. Otros autores señalan que en este siglo ha ocurrido el "segundo diluvio", el diluvio de la información, refiriéndose al crecimiento exponencial, explosivo y caótico de la cantidad de información que es accesible al individuo, la cual ha seguido multiplicándose en forma sostenida a través del tiempo. En este escenario una nueva cultura ha surgido,

para la cual el desarrollo de tecnologías digitales de información y comunicaciones ha sido vital, permitiendo que surjan nuevas condiciones y oportunidades para el desarrollo de los individuos y de la sociedad. El impacto tanto social como cultural, ha dado a la tecnología un rol preponderante dentro de la sociedad.

La tecnología, más que ser una entidad real que existe de manera aislada, es un punto de vista que enfatiza los componentes materiales y artificiales del fenómeno humano. Por tanto, es imposible separar lo humano de su entorno material. La tecnología es producida por una cultura, y una sociedad es condicionada por esta tecnología [Levy2001]. La tecnología provee acceso a ciertas posibilidades, cuyas repercusiones sociales podrían no existir sin su presencia.

Un ejemplo de este nuevo paradigma socio-tecnológico es el desarrollo que han tenido las tecnologías basadas en la Web, las cuales han posibilitado la masiva proliferación de entidades sociales soportadas por dicha tecnología, como por ejemplo las comunidades virtuales, o redes sociales en línea. Las personas han dejado de ser meros espectadores del contenido estático que predominaba la Web hace varios años atrás, tomando en la actualidad un rol mucho más participativo, opinando, colaborando mutuamente e incluso generando contenido en forma conjunta. La tecnología ha ofrecido un medio para la interacción entre las personas, a pesar de encontrarse ubicadas geográficamente en lugares alejados, y les ha permitido crear un sentido de pertenencia e identidad, a pesar de no existir necesariamente un contacto cara-a-cara. Como resultado, la sociedad ha cambiado, puesto que el surgimiento de estructuras sociales no está condicionado a la presencia física simultanea de las personas en un mismo lugar, sino que la tecnología ofrece un mecanismo para la formación virtual de dichas estructuras sociales. De esta manera, la tecnología cumple un rol fundamental, siendo el mecanismo que permite mediar las interacciones que ocurren dentro de estas nuevas estructuras sociales, dado que a través de ellas sus miembros desarrollarán las

características necesarias para su existencia.

Sin embargo, aplicar tecnología en este nuevo escenario no es una tarea trivial. Como se mencionó anteriormente, ya no se trata de ofrecer sistemas que se comuniquen entre sí, o sistemas utilizados por usuarios en forma aislada, como los procesadores de texto. Es necesario ofrecer el soporte adecuado a estas estructuras sociales virtuales, de tal forma que sea posible mediar y facilitar la interacción entre sus miembros. De esta forma, surgen diversas interrogantes: ¿cómo la tecnología puede apoyar eficientemente estas nuevas estructuras sociales?, ¿cómo escoger la alternativa técnica adecuada? o ¿cómo determinar si una estructura social utiliza la tecnología adecuada?

Encontrar respuestas a estas interrogantes no es simple, no sólo por que la tecnología es cambiante, sino además debido a la propia naturaleza evolutiva de estas estructuras sociales, lo cual provoca que sus necesidades cambien a través del tiempo. Las decisiones que se tomen a lo largo de la vida de la comunidad virtual, u otro tipo de estructura social, serán fundamentales para asegurar su correcto crecimiento. Wenger [Wenger2002] señala que en general una comunidad virtual no se crea, sino que se cultiva, de forma orgánica como una planta. Para su crecimiento, es necesario proveer de todos los cuidados necesarios, dándole todos los elementos que requiera, y protegiéndola de todos los peligros existentes desde el interior y el exterior de la comunidad de usuarios de la red social. Esto implica una constante preocupación por el crecimiento adecuado de la comunidad para que sea lo más saludable posible. En una comunidad virtual este aspecto toma mayor importancia, puesto que el desconocimiento de las herramientas o estrategias adecuadas puede significar tomar decisiones equivocadas en el momento de crear la infraestructura informática que soporta una estructura social de este tipo.

El fenómeno anterior se vuelve más complejo si consideramos el hecho de que muchas de las comunidades existentes son administradas por sus mismos creadores. Ellos, por lo general, no siguen una estrategia específica para guiar el crecimiento y evolución de la comunidad, ni tampoco cuentan con las

herramientas adecuadas que les permitan tener una visión completa de lo que va sucediendo al interior de la comunidad. Por lo tanto, se ven imposibilitados para tomar las acciones necesarias para satisfacer las necesidades cambiantes que se van generando al interior de la comunidad virtual. De esta forma, los administradores pasan a ser meros espectadores de los acontecimientos que suceden en la comunidad, tanto los relacionados con el crecimiento de la misma, como de su posterior extinción.

¿Por qué sucede el escenario anterior? Principalmente por el hecho de que cualquier persona puede comenzar una comunidad virtual, ya sea utilizando plataformas existentes o implementando una propia. El creador de la comunidad virtual en algún momento se verá enfrentado al problema de tener que tomar decisiones que favorezcan el desarrollo de la comunidad, sin embargo no se encontrará necesariamente preparado para ello. Dado que la tecnología condiciona las interacciones entre las personas, una mala decisión que se tome limitará las posibilidades de la comunidad virtual que se desea apoyar. Una consecuencia directa de esto es que sus miembros no puedan interactuar de forma adecuada, lo cual provocará la desaparición paulatina de dicha comunidad.

El desafío planteado en este escenario se debe a que la solución no es meramente tecnológica, dado que para el correcto crecimiento de una comunidad virtual, u otras estructuras sociales, es fundamental considerar el rol que tiene el uso de la tecnología en su ciclo de vida. Para ello es importante contar con herramientas o metodologías que permitan comprender cómo las aplicaciones computacionales, que median las interacciones entre los miembros de una comunidad, facilitan o no el desarrollo mismo de la comunidad. Las técnicas software que no consideren los aspectos sociales de estas estructuras no podrán ser aplicadas satisfactoriamente, dado que este tipo de sistemas son utilizados por estructuras sociales.

1.2. Motivación

A pesar de que varias teorías reconocidas en el área de sociología identifican los diversos aspectos sociales que caracterizan a una comunidad virtual [Kim2000] [Wenger2002] [Henri2003] 00[Preece2001] [Preece2004] [Wellman1997], y de que se reconoce la necesidad de un modelo en forma de un proceso cíclico que permita comprender y explicar cómo la tecnología es realmente utilizada por una estructura social [Wenger2005], poco se ha hecho desde el punto de vista computacional para desarrollar métodos de evaluación de dichos aspectos sociales, en escenarios donde un sistema de software actúa como infraestructura mediadora de las interacciones sociales entre usuarios.

La mayor parte de los trabajos publicados se enfocan en evaluar aspectos basados en la representación de un grafo de la estructura social. De esta forma, se concentran en la detección de nodos de influencia, comunidades, conectividad, y otras características puramente estructurales. El problema de este enfoque es que omiten la dinámica existente al interior de la comunidad, y al hecho de que las relaciones sociales que se forman tienen un motivo, dentro del contexto de la comunidad, ya sea por la existencia de un propósito común o de la aplicación de las políticas de moderación de la comunidad. Cada miembro, al interactuar con otro, le da un significado a su participación no solamente dirigiendo su mensaje, sino que también a través del contenido específico que genera.

Para subsanar lo anterior, es posible ver algunos trabajos que incorporan técnicas de *text mining* para sus estudios [Lipizzi2016]. Sin embargo, dado que es una aproximación emergente, existen escasos trabajos que complementen el análisis estructural más tradicional mediante la inclusión de un análisis semántico basado en el contenido de las interacciones, que permita comprender cómo evoluciona una comunidad virtual y cómo los diferentes roles de los usuarios se van complementando durante su existencia.

La construcción paulatina de una identidad compartida, como resultado de un constante proceso de negociación de significado, es la base para la

existencia de una comunidad virtual saludable. Sin embargo cada comunidad construye un dominio propio, el cual está en constante evolución y no depende solamente de sus miembros, sino de la forma en la cual han interactuado.

1.3. Hipótesis y Objetivos de la Tesis

La hipótesis de trabajo en esta Tesis es la siguiente:

H1: La información semántica permite elaborar modelos de redes sociales más ajustadas a la realidad de la interacción social efectiva.

El objetivo general de este trabajo de tesis es proponer nuevas técnicas que permitan construir modelos de redes sociales de mayor ajuste a la realidad, para poder describir el funcionamiento de dichas estructuras sociales en base a la interacción social efectiva.

Los objetivos específicos que se desprenden del objetivo general son los siguientes:

1. Complementar el análisis estructural tradicional, basado en la representación como grafo de la estructura social, con la utilización de análisis semántico del contenido de las interacciones (mensajes) entre los miembros de una comunidad virtual.

2. Identificar los diversos roles de los usuarios en la vida de una comunidad virtual, mediante el análisis estructural y el análisis semántico de las comunicaciones entre usuarios de la estructura social.

3. Aplicar la estrategia propuesta en comunidades virtuales reales, para evaluar la utilidad y veracidad de los resultados obtenidos. Particularmente, como demostradores, se aplican la técnicas propuestas sobre dos comunidades virtuales en las cuales existe por lo menos un administrador que conoce la evolución y funcionamiento de la comunidad virtual,

de tal forma que podemos utilizarse su conocimiento para evaluar los resultados obtenidos.

1.4. Metodología Utilizada

En el trabajo de la Tesis se ha utilizado como base el proceso *Cross Industry Standard Process for Data Mining* (CRISP-DM) [Chapman2000] que provee un marco de trabajo para analizar la información generada por un sistema de software. Las actividades consideradas en esta Tesis son las siguientes:

1. Revisión del estado del arte relacionado con el análisis de aspectos sociales desarrollados en comunidades virtuales. El énfasis principal de esta revisión está en analizar los aspectos sociales relacionados con el concepto de sociabilidad y las técnicas de evaluación existentes.

2. Recolección de datos históricos desde las comunidades virtuales en estudio (después de que los administradores de las redes eliminaran todo rastro de información personal), utilizando una estructura de datos común para representar la información de la comunidad.

3. Análisis semántico de las interacciones de la comunidad, en base al uso de modelos de semántica latente.

4. Análisis estructural de la comunidad, en base al uso de de técnicas de análisis de propagación de influencia.

5. Evaluación de los resultados obtenidos.

1.5. Casos reales utilizados para la experimentación

Como demostradores en esta Tesis se utilizaron dos comunidades de práctica virtuales, tanto para la recolección de información histórica, como para realizar los experimentos que permitieron validar la utilidad y veracidad de los resultados obtenidos. Las comunidades a las que se hace referencia son las siguientes:

- *Comunidad Virtual I*: Se encuentra formada por una red social de personas que comparten el interés sobre la construcción de efectos de música, amplificadores y equipos de audio. Los miembros de la red social se caracterizan por estar orientados hacia "hazlo tú mismo". Se cuenta con datos de la comunidad entre los años 2005 y 2015, existiendo aproximadamente 2.500 miembros al momento de extracción de los datos, los cuales utilizan un sistema de foro de discusión. En sus más de 10 años de experiencia, sus miembros han compartido y discutido su conocimiento en torno a la construcción de sus propios plexis y efectos. Sin embargo, han aparecido nuevos temas, tales como la lutería, audio profesional, compra/venta de partes, etc. En un comienzo, la administración era una tarea más bien sencilla, efectuada por un solo miembro; sin embargo en la actualidad dicha tarea es realizada por varias personas debido al tamaño de la comunidad. La visión de los administradores y miembros expertos acerca de la comunidad, está basada mayormente en la experiencia y el tiempo durante el cual han participado en la comunidad. Ellos poseen algunas medidas globales básicas que les permiten comprender el estado de la comunidad, como por ejemplo, el número total de publicaciones, y miembros conectados, entre otros. Sin embargo, ellos no poseen información acerca el comportamiento de los usuarios, calidad de las publicaciones de estas personas, y cómo ellos contribuyen a los objetivos de la comunidad. Para realizar este trabajo

se tuvo acceso a la base de datos del sistema de foro de discusión, acceso a los logs del servidor Web, se pudo hacer encuestas a los miembros, y entrevistas a los administradores de la comunidad.

- *Comunidad Virtual II*: Esta comunidad se encuentra dentro de una empresa, y fue creada inicialmente con el objetivo de promover el intercambio de ideas para apoyar la innovación en dicha empresa. La comunidad cuenta con más de 500 miembros en sus más de 3 años de experiencia, los cuales utilizan un sistema Web desarrollado especialmente para la comunidad. En dicho sistema sus miembros pueden subir sus ideas de innovación, y el resto de los miembros puede opinar al respecto. La tarea de administración es realizada por un empleado contratado especialmente para dicha actividad, el cual sólo utiliza medidas globales básicas, como por ejemplo, el número total de publicaciones, y número de comentarios, entre otros. Se espera que el resto de la organización utilice el sistema (se agregarían 1.500 miembros más), sin embargo, no cuentan con indicadores que les permitan gestionar la comunidad de forma apropiada. Para este trabajo se tuvo acceso a la base de datos del sistema de software utilizado, a los logs del servidor Web de los últimos meses, se contó con la posibilidad de realizar encuestas a los miembros de la comunidad, y entrevistas al administrador.

1.6. Contribuciones de la Tesis

En el marco de la resolución del problema planteado, este trabajo de tesis contribuye con:

(a) un análisis, basado en la revisión del estado del arte, de los aspectos sociales existentes y sus estrategias actuales de medición,

(b) un método para describir el funcionamiento de una estructura social, mediante la combinación de análisis estructural y análisis semántico, y

(c) una guía respecto a cómo interpretar los resultados obtenidos de la evaluación.

Debido a que la mayor parte de los trabajos existentes, revisados en el Capítulo 2, se enfocan principalmente en analizar los aspectos que motivan la participación de los miembros de una comunidad, más que el análisis de la estructura social que emerge, se espera además que las soluciones que se proponen en el ámbito de este trabajo ayuden a desarrollar más este enfoque por parte de futuras iniciativas.

1.7. Publicaciones

Las siguientes publicaciones son el resultado directo del trabajo reportado en esta Tesis:

- Semantically enhanced network analysis for influencer identification in online social networks. Sebastián A. Ríos, Felipe Aguilera, JDavid Nuñez Gonzalez, Manuel Graña. Neurocomputing (online, 2017). DOI https://doi.org/10.1016/j.neucom.2017.01.123
- Leveraging social network analysis with topic models and the Semantic Web (extended). Sebastián A. Ríos, Felipe Aguilera, Francisco Bustos, Tope Omitola, Nigel Shadbolt. Journal of Web Intelligence and Agent Systems, 11 (4), p.p. 303-314 (2013).
- A Dissimilarity Measure for Automate Moderation in Online Social Networks. Sebastián A. Ríos, Roberto Silva, Felipe Aguilera. Proceedings of 4th International Workshop on Web Intelligence & Communities (WWW 2012), Lyon, France (2012).
- Leveraging Social Network Analysis with Topic Models and the Semantic Web. Sebastián A. Ríos, Felipe Aguilera, Francisco Bustos, Tope

Omitola, Nigel Shadbolt. Proceedings of Web Intelligence and Intelligent Agent Technology (WI-IAT) at the IEEE/WIC/ACM International Conference on Web Intelligence (WI 2011), Lyon, France, p.p. 339-342 (2011).

- Topic-based social network analysis for virtual communities of interests in the dark web. Gaston L'Huillier, Héctor Álvarez, Sebastián A. Ríos, Felipe Aguilera. SIGKDD Explorations, 12 (2), p.p. 66-73 (2011).

- Web Intelligence on the Social Web. Sebastián A. Ríos, Felipe Aguilera. Chapter of Book Advanced Techniques in Web Intelligence, Springer Verlag, p.p 225-249 (2010).

- Enhancing Social Network Analysis with a Concept-Based Text Mining Approach to Discover Key Members on a Virtual Community of Practice. Héctor Álvarez, Sebastián A. Ríos, Felipe Aguilera, Eduardo Merlo, Luis A. Guerrero. Proceedings of 14th International Conference on Knowledge-Based and Intelligent Information and Engineering System (KES 2010), p.p. 591-600 (2010).

- Topic-based social network analysis for virtual communities of interests in the dark web. Gaston L'Huillier, Héctor Álvarez, Sebastián A. Ríos, Felipe Aguilera. Proceedings of 10th ACM SIGKDD Workshop on Intelligence and Security Informatics (ISI-KDD 2010), Washington DC, USA, July 2010 (2010).

- Virtual Communities of Practice's Purpose Evolution Analysis using a Concept-based Mining Approach. Sebastián A. Ríos, Felipe Aguilera, Luís A. Guerrero. Proceedings of 13th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES 2009), LNAI 5712, Santiago, Chile, September 2009, p.p. 480-489 (2009).

1.8. Estructura de la Tesis

La estructura de la Tesis consta de los siguientes capítulos:

- **Capítulo 2**: Está dedicado a presentar el marco teórico y los trabajos relacionados, incluyendo un resumen de los aspectos sociales involucrados en los principales tipos de estructuras sociales, un resumen de las diferentes estrategias y técnicas de evaluación usadas en comunidades virtuales, y una descripción de los modelos de conceptos sociales existentes.

- **Capítulo 3**: Describe los casos de estudio usados como demostradores, que nos aportan los conjuntos de datos utilizados para los experimentos en este trabajo.

- **Capítulo 4**: Presenta en detalle las técnicas de análisis propuestas, con descripción de los aspectos metodológicos.

- **Capítulo 5**: Muestra los resultados de los experimentos realizados con los datos de las comunidades virtuales reales.

- **Capítulo 6** Finalmente se exponen las conclusiones de este trabajo y recomendaciones de trabajo futuro

Capítulo 2

Estado del arte

En este capitulo se presenta el marco teórico y los trabajos relacionados con esta tesis. Primeramente introducimos algunas definiciones de las estructuras sociales, como son las redes sociales, comunidades virtuales y comunidades de práctica. Seguidamente presentamos las estrategias de evaluación de las comunidades virtuales desde los puntos de vista sicológico y sociológico, para pasar a detallar las técnicas de evaluación ene redes sociales y comunidades virtuales. Finalmente proporcionamos una discusión del modelado de los aspectos sociales.

2.1. Estructuras Sociales

El acceso generalizado a servicios basados en Internet, i.e. los servicios Web, ha permitido que muchas personas puedan comunicarse e interactuar con otras sin importar su ubicación geográfica. Es posible conversar con otras personas, buscar a gente con intereses comunes, ayudar a otros en determinados problemas, compartir información, y participar en debates, etc. Estas actividades han provocado que el uso del computador pase de ser una actividad individual, a una actividad colectiva en la cual se crean diferentes nexos de interacción y cooperación con otras personas. De esta manera, In-

ternet ha permitido que emerjan nuevas estructuras sociales [Wellman1996] [Wellman2001] las cuales, a pesar de estar basadas en estructuras sociales ya existentes, poseen características propias [Kim2000] que deben ser consideradas por los estudios que tratan de analizarlas. Estas características propias son debidas al medio de interacción remoto que utilizan, el cual ya no es cara-a-cara, lo cual provoca que muchos de los rituales sociales que se practican en la interacción física en el mundo real [Wellman1999] no existan, estén limitados, o sean simplificados en el mundo virtual [Johnson2001]. Comprender cuáles son los principales aspectos de estas estructuras sociales es fundamental para poder evaluarlas adecuadamente. En este capítulo se presentan los aspectos sociales más importantes de estas estructuras que permiten comprender, tanto el contexto social en el cual se desarrollan estas estructuras sociales, como los aspectos que los sistemas de software deben incluir al momento de evaluar su funcionamiento.

2.1.1. Redes Sociales

Internet ha posibilitado no solamente conectar computadores a través de la red, sino que también personas [Breslin2007] [Wellman1997]. El uso del correo electrónico, foros de discusión, y otros sistemas han permitido que las personas puedan trabajar y colaborar en grupos en línea, facilitados por la formación de redes sociales.

Una red está constituida por un conjunto de objetos (llamados nodos) y un conjunto de relaciones entre estos objetos. Una red social es un conjunto de actores sociales relacionados a través de vínculos sociales, que pueden ser, por ejemplo, de amistad o trabajo cooperativo. Los actores sociales pueden ser personas individuales, grupos u otro tipo de organización social. Una red puede ser representada gráficamente a través de un sociograma (Figura 2.1). En dicha representación, se utilizan puntos para indicar a las personas y líneas para representar las relaciones, las cuales pueden ser sólidas o intermitentes, de acuerdo a si las relación es fuerte o débil.

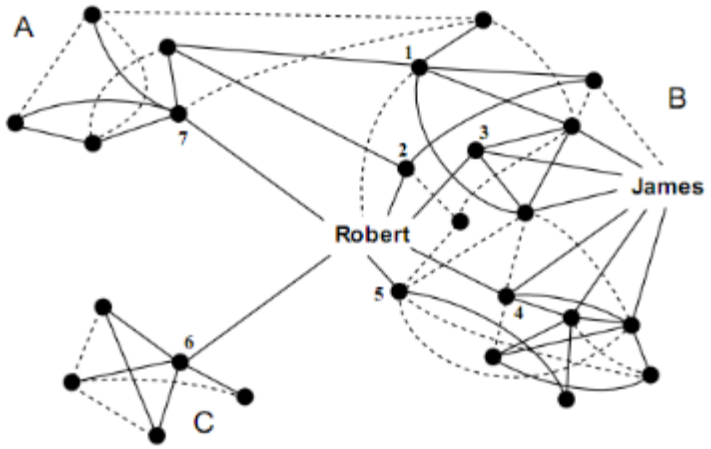

Figura 2.1: Representación gráfica de una red social en forma de un sociograma

Las redes sociales se han popularizado en los últimos años por la existencia de plataformas Web como Facebook o Twitter, entre otros, para la interacción en linea. En ellas, una persona puede comunicarse con otras, creando relaciones de tipo laboral, familiar, estudio, o amistad, entre otras. Las relaciones entre los usuarios de estas plataformas van conformando una gran red social global. En la red social, los siguientes aspectos son fundamentales:

1. Los actores y sus acciones son vistos como interdependientes, en vez de verlos como unidades independientes y autónomas.

2. Las relaciones existentes entre los actores son canales que permiten la transferencia o flujo de recursos (ya sean éstos tangibles o no) [Wellman1992].

Un grupo de personas que interactúa en línea construye implícitamente una red social, aunque ésta no sea definida explícitamente [Breslin2007]. Por lo tanto, dado que una persona puede participar en varios grupos, una misma persona puede ser parte de varias redes sociales.

Las diferencias entre una red social en línea con una red social física (que no existe en línea), surgen del mecanismo a través del cual interactúan dos actores dentro de la red: en uno será a través de un medio virtual y en otro será a través del mundo físico real. Gross [Gross2005] señala 3 grandes diferencias entre estas redes:

(I) mientras que en el mundo físico real, una relación social puede ser percibida por una persona de muy diversas formas, con grados de intensidad que van desde muy débil a muy fuerte, en una red social en línea tienden a ser reducidas a simples relaciones binarias formuladas como "es mi amigo o no lo es". Esta situación lleva al hecho de que una persona indique como su amigo a alguien a quien apenas conoce realmente [Boyd2004];

(II) mientras que la cantidad de nexos fuertes que una persona puede mantener en una red social en línea no aumenta significativamente con el uso de tecnología, el número de nexos débiles si puede incrementarse sustancialmente, dado que la comunicación en línea es más adecuada para ese tipo de nexos [Donath2004]. Finalmente,

(III) mientras que en una red social real una persona podría tener docenas de relaciones significativas y entre 1.000 y 1.700 "conocidos", en una red social en línea es posible encontrar cientos de "amigos" y cientos de miles de conocidos que se encuentran a no más de 3 grados de separación de la persona.

Lo anterior implica que una red social en línea es mucho más amplia y tiene muchas más conexiones débiles que una red social real [Gross2005]. Dicha situación provoca un desafío diferente, relacionado con la privacidad de las conexiones que posee una persona. Dado que una persona es mucho más "conocida", su información será también conocida por un mayor número de personas. Considerando lo anterior, ¿qué hace que una persona quiera hacer públicas sus conexiones con otras personas? Para Burt [Burt2002] el camino

más básico para establecer confianza en nuevas relaciones con personas, es que éstas ya sean conocidas por gente en las que uno ya confía. De esta forma, las conexiones que posee una persona constituyen una forma de desplegar y aumentar su *capital social* [Donath2004], el cual es reconocido por otras personas de la red social.

El término *capital social* se refiere a la idea generalmente aceptada de que participar en redes sociales u otras estructuras sociales tiene consecuencias positivas para un individuo [Portes1998]. Dicho capital es intangible y es un mecanismo a través del cual las personas adquieren conocimiento [Adler2002]. Putnam señala [Putman1996][Putman2000], basado en la evidencia, que existen tres formas de capital social [Wellman2001]:

1. Capital de Red: compuesto por las relaciones que posee una persona, con sus amigos, vecinos, parientes, compañeros de trabajos, etc., y todas aquellas relaciones que proveen algún beneficio, como compañía, ayuda emocional, bienes o servicios y sentido de pertenencia [Wellman2001];

2. Capital Participativo: la participación en organizaciones, de forma voluntaria, ofrece la oportunidad a las persona de crear logros en conjunto, articulando sus demandas y deseos, y

3. Compromiso Comunitario: el capital social consiste en mucho más que interacciones entre personas y participación en una organización social.

Las personas que tienen una actitud fuertemente positiva hacia una comunidad movilizarán su capital social de buena gana y de forma efectiva. Por lo tanto, conociendo las conexiones de un usuario, no sólo es posible determinar información de esa persona relacionada con sus amistades, gustos (musicales, artísticos, etc.) u otra información similar, sino que también información relacionada a cuán importante es la persona dentro de la red social, i.e. cual es su rol. Ahora bien, ser miembro de una red social tiene otras ventajas, tales como:

(a) permite conocer las necesidades actuales de los demás y estimula estas relaciones a través de contactos más frecuentes en el tiempo [Caspar2010];

(b) permite intercambiar información personal en forma de fotos, canciones y otros tipos de archivos; y

(c) permite conocer personas.

Es posible caracterizar una red social en línea a través del análisis de las diversas propiedades que posee, por ejemplo, a través de sus niveles de reciprocidad, densidad y formación de componentes a través del tiempo [Kumar2006]. Estudios previos [Kumar2006] [Mislove2007] [Zhu2010] han mostrado que grandes redes sociales en línea, como Flickr, Youtube, LiveJournal, Yahoo! 360 y Orkut, a pesar de ser diferentes, poseen valores similares de algunas propiedades mencionadas anteriormente, tales como la reciprocidad (entre un 62 % y 100 %). Este hecho muestra un hecho interesante, relacionado con el crecimiento similar de las redes sociales en línea, y la similitud en la forma en que las personas interactúan en las redes sociales.

2.1.2. Comunidades Virtuales

Buscar una definición precisa de comunidad no es una tarea sencilla. Es posible encontrar numerosas definiciones desde una perspectiva social, que han ido cambiando y redefiniendo en el tiempo [Barab2003] [Wellman1982]. Sin embargo, existen indicadores de comunidad que han sido adoptados por diversos investigadores [Jones1997][Rheingold1993][Wellman2000]:

- personas compartiendo algún interés común, experiencias y/o necesidades,
- enlazados por relaciones sociales a través de las cuales se obtienen recursos importantes,

- desarrollan fuertes sentimientos interpersonales de pertenencia y necesidad mutua, y
- se da el surgimiento de un sentido de identidad compartida.

Los tipos de comunidades que se crean a través de Internet difieren en varios aspectos respecto al concepto tradicional de comunidad, principalmente por el hecho de que son mediadas a través de sistemas de comunicación basados en computadores. A estos tipos de comunidades se les denomina "comunidades virtuales" o "comunidades en linea". En una comunidad virtual las personas pueden o no encontrarse cara-a-cara con otras, y el intercambio de mensajes e ideas se efectúa a través de la red [Rheingold1993]. Una comunidad virtual existe y juega un rol socializador de la misma forma que lo hacen las comunidades "reales" [Wenger2005] [Rheingold1993].

Las comunidades virtuales pueden ser vistas como una red social, dado que la red de computadores que conecta a las personas permite crear enlaces entre las personas con un significado social [Wellman1997]. De la misma forma, un grupo de trabajo también es una red social, siendo la principal diferencia con una comunidad virtual que las relaciones entre las personas se encuentran delimitadas firmemente en su frontera (no existen relaciones con personas que no pertenecen al grupo de trabajo) y son muy densos (cada persona se encuentra relacionada con la mayoría de los miembros del grupo) [Wellman1997]. Una comunidad virtual se crea a partir de la continuidad de las relaciones entre sus miembros, pero es experimentada a través de actividades específicas localizadas en el tiempo y el espacio [Wenger2005]. Los principales tipos de comunidades que se estudian en la actualidad son los siguientes:

1. Comunidades de interés [Marathe1999]. Son aquellas comunidades en las cuales sus integrantes comparten el mismo interés en algún tópico (y por tanto todos ellos poseen un *background* común). Ejemplos de

este tipo de comunidad son: los fan club de grupos musicales, los grupos de personas interesadas en los planetas del sistema solar, etc. Otro tipo de comunidades son las llamadas comunidades de pasión [Carotenuto1999], las cuales poseen una definición muy similar.

2. Comunidades de práctica [Wenger1999][Wenger2005][Shummer2004]. Son aquellas comunidades en las cuales sus integrantes comparten una misma profesión o interés por una actividad específica. Generalmente sus miembros se encuentran fuertemente involucrados en la comunidad. Ejemplos de este tipo de comunidad son: la comunidad de programadores de Java, la comunidad *Open Source*, comunidades dentro de empresas, etc.

3. Comunidades de propósito [Carotenuto1999]. Son aquellas comunidades en las cuales sus miembros comparten el mismo objetivo (de corto plazo generalmente). Ejemplos de este tipo de comunidad son los compradores de una librería virtual, los cuales comparten el objetivo de encontrar y comprar un libro. Los miembros tienen un propósito fundacional que es desaparece una vez que el objetivo es alcanzado. Generalmente los miembros de este tipo de comunidad no realizan actividades que excedan los propósitos de la comunidad ni comparten necesariamente el mismo interés [Carotenuto1999].

Existen otros tipos de comunidades que no se detallan en este trabajo. Una revisión de éstas puede ser encontrada en [Barab2003][Henri2003].

Basados en la teoría de Wenger sobre conocimiento [Wenger2002], Henry & Pudelko [Henri2003] establecen una relación entre los distintos tipos de comunidades. Existe una relación entre el tipo de comunidad que se forma y la cohesión e intencionalidad de la comunidad misma (Figura 2.2). Por ejemplo, en una comunidad de interés se presenta la menor cohesión posible y la intencionalidad, al momento de la creación de la comunidad, es leve. En el otro extremo, en cambio, encontramos a las comunidades de práctica, en

Figura 2.2: Tipos de comunidades virtuales (Figura obtenida de Henri2003).

las cuales el grupo de personas que la componen se encuentran fuertemente cohesionadas, y en donde la formación de la comunidad es altamente intencionada. A continuación describimos las comunidades de práctica que son el objeto de estudio de esta Tesis.

2.1.3. Comunidades Virtuales de Práctica

Las Comunidades Virtuales de Práctica (VCoP por su sigla en inglés) han experimentado un crecimiento explosivo en los últimos años. El valor de este tipo de comunidades es que permiten establecer relaciones entre personas que quieren compartir o aprender acerca de un tema específico, basados en la interacción de dichos miembros [Wenger2002]. El rol de Internet en su desarrollo es fundamental, puesto que facilita la interacción entre los miembros de la comunidad sin necesidad de contacto presencial, como requiere una comunidad "real". Para poder soportar este tipo de interacciones la Web ofrece un medio de interacción mucho más dinámico en comparación con otras tecnologías como el correo o la mensajería instantánea. Las herramientas más comunes utilizadas en la Web son los foros, wikis, y otras herramientas similares.

Para el funcionamiento de una VCoP es muy importante poder generar, almacenar y mantener el conocimiento resultante de la interacción entre sus miembros. El éxito de una VCoP depende en gran parte de los mecanismos de administración [Probst2008] y de las participación de los "miembros claves" de la comunidad (también llamados "líderes" [Bourhis2005] o "miembros principales" [Probst2008]). Del mismo modo, el objetivo de cada miembro de la comunidad virtual de práctica es aprender adquiriendo cierto conocimiento que poseen otros miembros de la comunidad. Por lo tanto, el contenido semántico de las interacciones es un aspecto fundamental que debe ser considerado en el estudio de las VCoP.

Plaskoff [Plaskoff2003], basado en su propia experiencia en una gran compañía farmacéutica, ha sugerido que para que exista una VCoP deben darse 3 actitudes de los miembros de la comunidad, de los cuales depende cultivar con éxito una VCoP dentro una organización: *believing*, *behaving*, y *belonging*. *Believing* se refiere a la idea de que los miembros deben creer en el valor intrínseco de la comunidad. *Behaving* indica que los miembros definen, crean y siguen normas de comunidad. *Belonging* significa que cada miembro cultiva un sentido de pertenencia con la comunidad.

Por otro lado, para que una VCoP pueda funcionar a través de Internet, primero es necesario crear una comunidad virtual. Sin embargo, la creación de una comunidad virtual no garantiza, necesariamente, que una VCoP llegue a ser desarrollada, puesto que, además, es preciso que exista un esquema de aprendizaje basado en tareas [Johnson2001]. Existen 3 elementos constitutivos que hacen que una estructura social pueda ser considerada como una "comunidad de práctica" (CoP) y no sólo como una "comunidad": dominio, comunidad y práctica [Wenger1999]:

1. Dominio. Una comunidad de práctica no es sólo un club de amigos o una red de conexiones entre personas. Tiene una identidad definida por un dominio compartido de intereses. Ser miembro de la comunidad implica compromiso, y por tanto una competencia (responsabilidad) comparti-

da que distingue a los miembros de otras personas [Wenger1999].

2. Comunidad. Al perseguir sus intereses en el dominio de la comunidad, los miembros se interesan en participar en actividades y discusiones, ayudar a otros y compartir información. Construyen relaciones que les permiten aprender de los demás. Es necesario ïnteractuarz .ªprender de otros"para obtener una comunidad de práctica [Wenger1999].

3. Práctica. Una CoP no es sólo una comunidad de personas con intereses quienes tienen un gusto en común. Los miembros de una CoP son prácticos, ellos desarrollan un repertorio compartido de recursos: experiencias, historias, herramientas, formas de direccionar problemas recurrentes, etc. El desarrollo de recursos y prácticas compartidas puede ser un proceso intencionado o un subproducto de la actividad de la comunidad.

Un aspecto fundamental de las comunidades de práctica es su evolución temporal [Wenger2002]. La vida de una comunidad es un viaje de descubrimiento que le ayuda a reinventarse continuamente [Wenger2005] [Preece2004] [Mynatt1997]. En la evolución de una comunidad es posible identificar 5 etapas (ilustradas en la Figura 2.3) [Wenger2002]:

1. Descubrimiento, los usuarios descubren la necesidad de la red para comunicarse con sus semejantes.

2. Incubación, comienza la creación de la infraestructura y los primeros usuarios se unen y trabajan conjuntamente.

3. Expansión, la comunidad comienza a ser conocida y atrae nuevos usuarios que buscan adquirir los conocimientos y recursos de la red.

4. Preservación, se mantiene estable la comunidad con una actividad similar a la inicialmente propuesta y en contacto con el exterior.

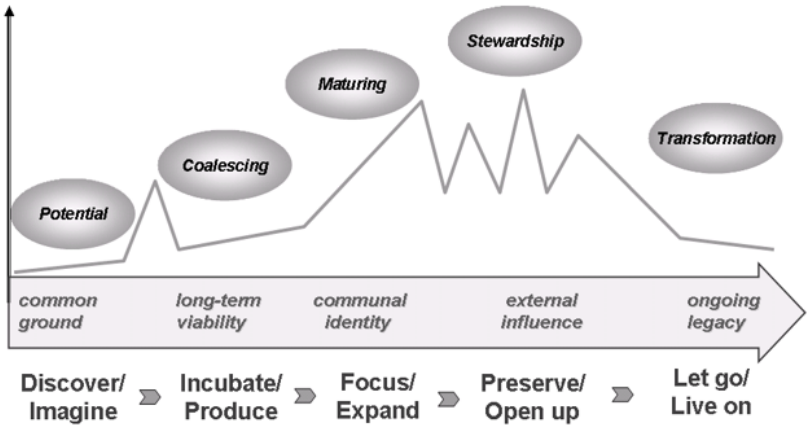

Figura 2.3: Evolución de una comunidad de práctica (Fig. obtenida de [Wenger2002]).

5. Transformación, los contenidos de las comunicaciones en la red se transforman y como consecuencia la red pierde interés como comunidad de práctica y se convierte en otra cosa.

De la misma forma, la modalidad de participación de los miembros de una comunidad puede evolucionar con el tiempo, desde observadores, hasta participantes ocasionales o líderes y coordinadores de la comunidad. Wenger define una serie de niveles de participación de los miembros de una comunidad [Wenger2002], a partir de las relaciones que forjan con los demás miembros (ilustradas en Figura 2.4).

Como se señaló anteriormente, una VCoP comparte características con las comunidades virtuales y redes sociales en línea, dado que éstas corresponden a estructuras mucho más genéricas en las cuales se basa la definición de una VCoP. Este hecho provoca que el análisis de este tipo de estructuras sociales sea más complejo que el de otras estructuras más genéricas (redes sociales por ejemplo), dado que será necesario estudiar tanto los aspectos que comparte con estas estructuras sociales, así como sus aspectos propios. En la Tabla 2.1, se muestra un resumen de los aspectos sociales más importantes involucrados

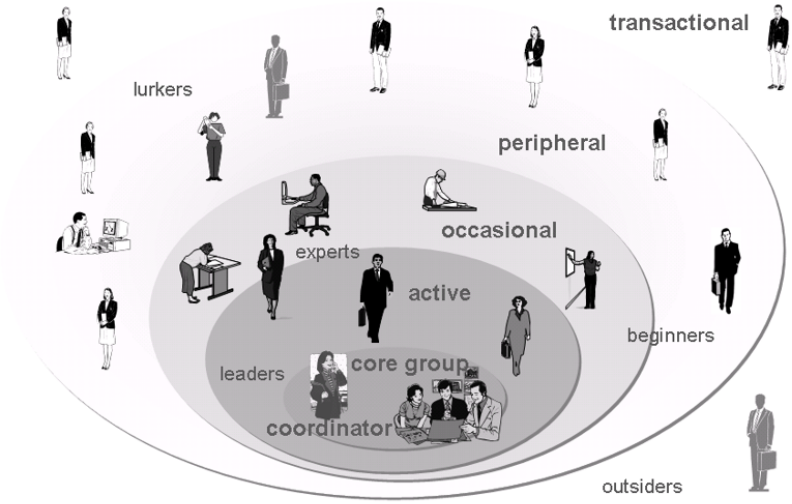

Figura 2.4: Niveles de participación de los miembros de una comunidad (Figura obtenida de [Wenger2002]).

en cada una de las estructuras sociales mencionadas anteriormente, las cuales sirven como una guía para el análisis de una comunidad virtual de práctica. El detalle de los aspectos sociales es posible revisarlo en el Anexo B.

Cuadro 2.1: Resumen de los aspectos sociales de las estructuras sociales en estudio, desde la perspectiva de diversos autores.

Autor	Aspecto Social	Descripción
Redes Sociales		
Garton [44]	Actores	Se refiere a las entidades sociales que son parte de la red, por ejemplo, personas, organizaciones, etc.
	Relaciones	Se refiere a las conexiones existentes entre los diferentes actores. Una relación es caracterizada por su contenido, dirección y fuerza.
	Lazos	Un lazo conecta a un par de actores a través de 1 o más relaciones. También se pueden describir en base a su contenido, dirección y fuerza.
	Multiplicidad	Entre más relaciones posee un lazo, entonces más multiplicidad hay en el lazo.
	Composición	La composición de una relación o lazo es derivada de los atributos sociales de ambos participantes.
Wellman [7]	Rango	Se refiere al tamaño y heterogeneidad de la red social.
	Centralidad	Se refiere a quién es central en la red social.

	Roles	Se refiere a la similitud en el comportamiento de los miembros de la red.
	Grupos	Es una estructura que se descubre en forma empírica, a través del análisis de los patrones de relaciones entre los miembros de la población. Los grupos emergen como conjuntos altamente conectados de actores.
Comunidades Virtuales		
Preece [45]	Personas	Corresponde a las personas que interactúan con otras en la comunidad, quienes poseen necesidades individuales, sociales y organizacionales.
	Propósito	Una comunidad comparte un interés, necesidad, información, servicio, o soporte, el cual provee un motivo a sus miembros para pertenecer a la comunidad.
	Políticas	El lenguaje y protocolo que guían las interacciones entre las personas y contribuyen al desarrollo del "folclore" o rituales que dan un sentido de normas sociales aceptadas e históricas.

Henri & Pudelko [4]	Intención	También denominado objetivos de la comunidad.
	Inicio	Los métodos a través de los cuales se crea el grupo inicial que conforma la comunidad.
	Evolución	Se refiere a la evolución temporal, tanto de la intención como de los métodos que forman grupos al interior de la comunidad.
Kim [3]	Propósito	Se refiere a los objetivos de los miembros de la comunidad.
	Lugares	Se refiere a los lugares en los cuales los miembros trabajan en conjunto.
	Perfiles	Es un conjunto de información que dice algo acerca de quién es un miembro, en el contexto de la comunidad.
	Roles	Se refiere a la existencia de distintos tipos de miembros al interior de la comunidad.
	Liderazgo	Se refiere al más visible de los roles. Corresponde a aquellos miembros que ayudan a mantener la comunidad funcionando.
	Etiquette	Es un conjunto de comportamientos – o estándares dentro de la comunidad – que un grupo de personas acepta.

	Eventos	Se refiere a las reuniones al interior de la comunidad, las cuales ayudan a definir la comunidad, recordarles a sus miembros lo que tienen en común y de qué se trata la comunidad en la cual participan.
	Rituales	Corresponden a celebraciones de carácter social (por ejemplo el cumpleaños de uno de los miembros), los cuales permiten crear un sentido de identidad al interior de la comunidad.
	Subgrupos	Se refiere a grupos pequeños de personas, en los cuales sus miembros forman sus más profundas relaciones y fuertes lealtades.
Selznik [46], Schwier [47]	Historia	La existencia de historias y cultura compartidas son un elemento que fortalece los lazos comunitarios.
	Identidad	Las comunidades promueven un sentido de identidad compartida.
	Mutualidad	Las comunidades brotan y se mantienen por las interdependencias y reciprocidades existentes.

	Pluralidad	Las comunidades obtienen gran parte de su vitalidad al ser "asociaciones intermedias" entre familia, iglesias y otros grupos periféricos.
	Autonomía	A pesar de haber un énfasis en la identidad grupal, es importante respetar y proteger la identidad individual de cada miembro.
	Participación	La participación social en la comunidad, especialmente aquella que promueve la autodeterminación, apoya la autonomía y permite que la comunidad se sostenga.
	Integración	Todos los otros elementos dependen de las normas de apoyo, las creencias y prácticas.
Comunidades de Práctica		
Wenger [39]	Comunidad	Se refiere a las relaciones que edifican los miembros para poder aprender entre sí.
	Práctica	Se refiere a un repertorio compartido de recursos: experiencias, historias, herramientas, las maneras de abordar los problemas recurrentes.

	Dominio	Una CoP tiene una identidad definida por un dominio compartido de interés.
Plaskoff [42]	Believing	Se refiere a la idea de que los miembros necesitan creer en el valor intrínseco de la comunidad.
	Behaving	Indica que los miembros a desarrollar y seguir las normas de una comunidad.
	Belonging	Esto significa que los miembros de cultivar un sentimiento de pertenencia a una comunidad.
Lesser [48]	Conexiones	Existe una serie de conexiones entre los individuos de la comunidad. En otras palabras, los individuos perciben que ellos mismos son parte de una red (dimensión de estructura).
	Relaciones	Corresponde a un sentido de identidad que es desarrollado a través de las conexiones (dimensión de relaciones).
	Contexto Común	Los miembros tienen un interés común o un entendimiento compartido del ambiente en el que participan (su organización, por ejemplo) (dimensión cognitiva).
Saint-Onge [49]	Práctica	Se refiere a las actividades que realizan los miembros.

	Personas	Se refiere a quienes están involucrados en la comunidad.
	Capacidades	Se refiere a la capacidad de aprovechar las ventajas competitivas que poseen.

2.2. Estrategias de Evaluación de Comunidades Virtuales

Es posible analizar el comportamiento de una comunidad virtual desde diversos puntos de vista. Los más clásicos se orientan al análisis del contenido generado por los miembros de la comunidad al interactuar a través del tiempo, sin considerar que existe una estructura social que interactúa a través de un sistema de software. Por ejemplo, podemos evaluar la comunidad a través del análisis de la calidad de las respuestas que se dan dentro de una comunidad [Anderson2012], el análisis del comportamiento semántico de los comentarios generados por los miembros de una comunidad [Siersdorfer2014], o el análisis del sentimiento/opinión de un miembro de la comunidad acerca de un tema en particular [Kontopoulos2013]. El inconveniente de este punto de vista es que, al evaluar únicamente el resultado de la interacción entre los miembros de una comunidad virtual, se dejan de lado los aspectos sociales que se desarrollan al interactuar los miembros. Por lo tanto, se ignora la información de la estructura social que emerge como consecuencia de dichas interacciones.

Por otro lado, las aproximaciones más modernas sí consideran los aspectos de la estructura social que emerge como resultado de las interacciones entre las personas. Es posible distinguir dos puntos de vista: (a) el sicológico que enfatiza el estudio de los factores humanos que influyen en la participación de los usuarios en una estructura social en línea, y (b) el sociológico que se dirige

a analizar la estructura social que se forma producto de las interacciones entre los miembros de una comunidad, considerando a dicha estructura social como una entidad con sus propias características y formas de funcionamiento. A continuación se detallan estos dos enfoques, indicando ejemplos de trabajos que utilizan dichos enfoques.

2.2.1. El Punto de Vista Sicológico

Esta aproximación al problema de la evaluación de la red social en lina se basa en teorías tales como la teoría cognitiva social, la teoría de capital social, la teoría de redes sociales y la teoría del compromiso [Chiu2006]. Entre los trabajos relacionados que utilizan enfoques sicológicos se encuentran, por ejemplo, trabajos que se orientan en comprender los factores que motivan a una persona a continuar participando en una red social [Lin2011] o a compartir conocimientos específicos con otros miembros [Yan2016], o la reacción que ocasiona la participación de un miembro sobre otros, basados en la estimación del beneficio percibido por los miembros de una estructura social virtual al participar y la influencia que puede ocasionar la diferencia de género en el beneficio percibido [Zhou2014]. También hay trabajos basados en la teoría de balance estructural de Heider [Cartwright1956], al medir el beneficio percibido por otros usuario o grupos de usuarios [Danescu2009].

Existen otros trabajos que se concentran en el estudio de roles específicos que emergen dentro de la estructura social, como por ejemplo los observadores de una comunidad virtual [Sun2014] o los líderes y los tipos de liderazgo en el interior de una comunidad [Zhu2013], y la forma en que éstos influyen en la motivación y participación de otros miembros. Una de las ventajas de estos enfoques es que permiten realizar predicciones acerca del comportamiento futuro de los miembros de una comunidad. Por ejemplo, West et al. [West2014] predicen la opinión que tendrá un miembro A acerca de otro miembro B, basado en la mezcla de dos tipos de informaciones: la red social a la que pertenece la persona A y lo que escribe, esto es analizando lingüísticamente los

sentimientos que expresa. Por otra parte, Cheng et al. [Cheng2014] analizan cómo la evaluación entre miembros (a través del uso de técnicas de *rating* de los comentarios publicados) afecta el comportamiento futuro de un autor dentro de una comunidad, regulando la calidad y cantidad de contribuciones futuras de sus miembros.

La principal limitación de los trabajos anteriores es que no permiten evaluar el funcionamiento de una comunidad virtual como un todo, dado que su orientación es principalmente hacia el análisis del comportamiento de los individuos que participan dentro de una comunidad virtual o red social en línea.

2.2.2. El Punto de Vista Sociológico

La mayor parte de los trabajos que siguen esta estrategia están centrados en el uso de algoritmos de Análisis de Redes Sociales (SNA – Social Network Analysis), los cuales permiten obtener información acerca de la estructura de la comunidad a través del análisis automático de los datos que genera la estructura social. Por ejemplo, es posible encontrar trabajos que analizan la estructura de los usuarios utilizando las propiedades de desigualdad de la participación [Raeth2009], trabajos que buscan definir estrategias para detectar comunidades dentro de una red social [Fortunato2010][Kim2015], trabajos que realizan un análisis de las funciones relacionadas con la moderación [Gairín-Sallán2010][Matzat2014], o detectan miembros que cumples roles específicos, como el de intermediarios [Toral2010]. También, dentro de este enfoque, existen trabajos que analizan la dinámica de la estructura social y cómo se propaga la información en la red social. Por ejemplo, en [Myers2014] se analizan dos dinámicas: (1) creación y destrucción de la red subyacente de seguidores, y (2) la dinámica del flujo de información a través de la re-publicación de contenidos generados por otros miembros.

Lo que caracteriza a la mayoría de los trabajos antes descritos, es la utilización de técnicas que analizan los *logs* que contienen el historial de la

actividad de los miembros de estas estructuras sociales. Si bien hay un mayor énfasis en el uso de técnicas de análisis de redes sociales, es posible ver algunos trabajos que incorporan técnicas de *text mining* para sus estudios [Lipizzi2016]. Trabajos más frecuentes se enfocan en estudiar tipos específicos de comunidades virtuales [Euerby2014]. Sin embargo, éstos analizan solamente las conexiones entre los miembros de una comunidad, más que las propiedades que la diferencian de otras estructuras sociales [Dubé2003].

A pesar de que los trabajos anteriores utilizan información acerca de la estructura social que se genera por la actividad de la comunidad virtual, no obtienen indicadores globales que puedan ser utilizados como apoyo en la toma de decisiones por parte de los administradores. Tampoco consideran aspectos propios de las comunidades virtuales, sino que las analizan como si sólo fueran redes sociales en línea, con lo cual dejan de lado todos los aspectos que caracterizan a una comunidad virtual. Hay también otros trabajos intentan generar dichos indicadores, sin embargo las métricas que se obtienen no dan cuenta de los aspectos sociales que se desarrollan, sino que se enfocan en aspectos particulares del área de estudio. Por ejemplo, en marketing se usan este tipo de indicadores para medir la importancia de una marca [Hassan2014], o para mejorar las sugerencias de un sistema de recomendación [Li2014] [Lau2014]. También se usan para analizar la evolución de lo que siente un grupo de personas acerca de diversos tópicos de interés [Amer-Yahia2012], para comparar distintos foros utilizados por comunidades virtuales, y determinar así los tipos de miembros que participan [Jones2011].

2.3. Técnicas de Evaluación en Redes Sociales y Comunidades Virtuales

El proceso de evaluación es el único que puede garantizar de forma científica el efecto de las acciones de gestión que son tomadas en un momento dado. Existen diversas aproximaciones a la definición de la técnica de evaluación,

desde aquellas que evalúan al software en sí mismo, hasta aquellas que intentan evaluar la estructura social resultante. A continuación se detallan las principales técnicas de evaluación, de acuerdo a los principales aspectos examinados por cada aproximación.

2.3.1. Análisis de Redes Sociales

El área de Análisis de Redes Sociales (SNA de las siglas en inglés) tiene como objetivo estudiar estructuras sociales modeladas como redes, es decir, como un conjunto de actores y relaciones sociales entre estos actores. A diferencia de otros tipos de análisis, SNA no es lineal [Lave1991], es decir, su análisis no puede ser realizado incrementalmente, sino que debe ser realizado sobre la estructura completa [Rogers1981]. Esta propiedad de SNA se debe a que está basada en un análisis estructural [Wellman1998], es decir, el análisis está basado en las relaciones entre los actores y los patrones que surgen en la red producto de estas relaciones. Para realizar un análisis SNA es posible identificar 6 diferentes técnicas [Breiger2003], entre las cuales se destacan:

1. Métricas: El objetivo de utilizar métricas es medir propiedades propias de la red social, de sus actores o un conjunto de ellos. Por ejemplo, es posible medir la densidad de una red social para determinar cuán relacionados se encuentran sus miembros, o es posible medir la centralidad de un actor, para determinar su importancia dentro de la red social. Entre las métricas asociadas a los nodos de la red podemos encontrar [Lave1991] [Wasserman1994]:

    - Centralidad: El grado en el cual un actor se encuentra en un rol central dentro de la red.
    - Prestigio: También conocido como estatus, esta medida intenta cuantificar la confianza que posee un actor en particular respecto de un conjunto de actores.

- Homofilia: El grado en el cual actores similares en roles similares comparten información.
- Aislamiento: Un actor que no posee lazos con otros actores.
- Puerto: Un actor que conecta la red con las influencias externas (exterior).
- Punto de Corte: Un actor cuya remoción resulta en caminos inconexos dentro de la red.

2. Entre las métricas que son propias a una red a nivel gloval, podemos encontrar [Lave1991]:

   - Centralización: La fracción de actores principales dentro de la red.
   - Accesibilidad: El número de vínculos que conectan a los actores.
   - Conectividad: La capacidad de los actores para llegar de uno a otro recíprocamente, es decir, la capacidad de elegir una relación entre ambas partes.
   - Balance: La medida en que los lazos en la red son directos y recíprocos.

3. Subgrupos Cohesivos (comunidades): Se refiere a subconjuntos de actores cuyas relaciones son relativamente fuerte, directa, intensa, frecuente o positivas. Entre los métodos para detectarlos, podemos encontrar:

   - Completa Mutualidad: Este método se basa en la búsqueda de los subgrupos cohesivos en los cuales existen todas las relaciones sociales posibles entre todos los actores. Para ellos, se basa en la definición de un clique, es decir, un subgrafo maximal completamente conectado de al menos tres nodos.
   - Accesibilidad y Diámetro: Este método se basa en una extensión del método anterior. Estos grupos son importantes si sabemos que

los procesos sociales más importantes se producen a través de intermediarios. Se basa en el hecho de que no todos los miembros de un subgrupo se encuentran conectados directamente, sin embargo, las rutas que los conectan son relativamente cortas. Ellos se basan en la definición de n-clique (el cual es el subgrafo maximal de distancia geodésica más larga entre dos nodos cualquiera no mayor a n; una ruta geodésica incluye cualquier nodo dentro del grafo), n-clans (es un n-clique en el cual la distancia geodésica entre todos los nodos en el subgrafo no es mayor que n para los caminos del grafo) y n-clubs (el cual es un subgrafo maximal de diámetro n, es decir, la distancia entre todos los nodos del subgrafo es menor o igual a n)

- *Nodal Degree*: este enfoque está basado en las restricciones sobre el número mínimo de actores adyacentes a cada actor en un subgrupo. Está basado en los conceptos de k-plexos (un k-plexo es un subgrafo maximal en el cual cada nodo en el subgrafo le puede faltar no más de k lazos a otros miembros del subgrafo) y k-cores (un k-core es un subgrafo en el cual cada nodo es adyacente a por lo menos k nodos dentro del mismo subgrafo)
- Matrices de Permutación: El propósito de las matrices de permutación es, dada una red social, representarla como una matriz NxN, reordenando sus columnas y filas con el objetivo de identificar bloques al interior de la red social. Cada fila y columna n corresponde a un miembro, y cada celda indica la existencia o no de un relación social entre dos miembros de la red social [Xu2005].

4. Visualización: Una red social puede ser modelada como un grafo, representando los actores sociales por círculos y las relaciones por flechas que conectan esos círculos. Esta representación visual es denominada sociograma. Permiten explorar y entender una red social a través del análisis

visual de sus relaciones. La complejidad de un sociograma se encuentra en la técnica que se utiliza para dibujar la red social [Huang2006], entre las cuales es posible encontrar:

- Disposición Circular: Todos los nodos son colocados en un círculo [Scott2000].
- Disposición en Radio: Los nodos son colocados en círculos concéntricos, en el cual los nodos más centrales son ubicados en el centro del diagrama.
- Disposición Agrupada: Es utilizado para desplegar información acerca de grupos, mostrando los nodos cerca de aquellos que comparten alguna característica.
- Disposición Libre: en el cual no hay ningún criterio para el despliegue, solamente que sea legible.

2.3.2. Análisis de Comunidades Virtuales

De acuerdo con [Preece2003] podemos identificar 4 formas de analizar una comunidad virtual:

- Etnografía y técnicas asociadas: el propósito de la investigación etnográfica es estudiar un grupo desde el punto de vista de sus miembros/participantes. Es un método de investigación cualitativo para comprender cómo la tecnología es usada in-situ [Fetterman1998] [Preece2003]. Un ejemplo de esta técnica es la aplicación de un análisis etnográfico a una comunidad virtual que usa una herramienta IRC como medio de comunicación entre sus miembros [Nocera2002]. Dado que la etnografía, en su forma original, no contempla las dificultades propias de la participación asíncrona y el hecho de que los miembros son virtuales, diversos autores han propuesto variantes a la técnica original, tales como la netnografía [Kozinets2002] y la cyberetnografía [Robinson].

- Cuestionarios: son útiles para recolectar información demográfica y tienen la ventaja que pueden ser distribuidos por mano a los participantes locales, o enviados por correo electrónico, o estar en la Web [Preece2004] [Preece2003]. Sin embargo, a pesar de que los cuestionarios proveen información útil de los miembros de una comunidad [Koh2007], no son suficientes. Es recomendable el uso de una evaluación secundaria, cuando sea posible, para reducir la subjetividad de la opinión de los miembros de la comunidad.

- Experimentos y cuasi-experimentos: estos estudios son valiosos para evaluar la usabilidad de las interfaces y la reacción de los usuarios a las nuevas características de las interfaces de usuario. Esta aproximación se usa habitualmente para investigar el impacto de cambios en el diseño del software que utiliza una comunidad en línea [Preece2004] [Sudweeks1999]. Para aplicar con éxito esta modalidad de evaluación es necesario que los miembros de una comunidad virtual interactúen en un ambiente controlado.

- Minería de Datos y Análisis de Redes Sociales: Este tipo de análisis consiste en utilizar los registros (*logs*) generados por el sistema de gestión de la red social para descubrir información útil de la comunidad virtual. El interés principal es el estudio de la naturaleza social de la comunidad, es decir, el estudio de sus miembros y las relaciones que ellos establecen [Preece2004]. En estos estudios típicamente se aplican dos clases estrategias. Las primeras son aquellas basadas en la visualización de la red social subyacente. Como resultado, es posible deducir o establecer la existencia de patrones de comportamiento u otras estructuras sociales dentro de la red social [Arenas2004]. Segundo, aquellos que intentan cuantificar de alguna manera las relaciones que se crean dentro de la comunidad [Ehrlich2007]. Para ello, es común definir y medir un conjunto de métricas e indicadores que permiten monitorizar

el comportamiento de la comunidad.

2.4. Modelado de Aspectos Sociales

En [Aguilera2017] se presenta una guía para el análisis de los aspectos sociales desarrollados por una comunidad, basándose en el concepto de sociabilidad; específicamente trata de tres dimensiones en las que se desarrolla la comunidad virtual que identificamos como **Personas**, **Propósito** y **Políticas** [5], que se describen a continuación:

- **Personas** [5][44][49]: Las personas son una parte fundamental de una comunidad virtual, pues son ellas las que la conforman. Sin personas no existen las comunidades. Las discusiones, la generación de nuevas ideas, y el constante intercambio de los contenidos son los que distinguen a una comunidad de las páginas Web tradicionales. Una persona es mucho más que un usuario de un sistema. Una persona es una entidad social en si misma: tiene sus propios objetivos y metas. Las personas son las que se relacionan entre sí para conformar una nueva estructura social. Para un mejor entendimiento, la dimensión **personas** la hemos dividido en los siguientes indicadores, que son los usados por los administradores para entender los que está pasando con la comunidad desde esa perspectiva:

  - Niveles de Participación [96][46]: En una comunidad virtual, no todos sus miembros participan de la misma forma. Esto se debe a la voluntariedad de la participación, que provoca una constante evolución, en la cual los miembros van cambiando de forma autónoma su forma de participar a lo largo del tiempo. Por ejemplo, es posible encontrar moderadores (quienes guían las discusiones), profesionales (quienes dan opiniones) y observadores (quienes silenciosamente observan el funcionamiento de la comunidad). Por

tanto, es esperable que existan miembros que participen más activamente y otros que participen menos, o incluso sólo observen en forma pasiva. Específicamente se espera, en la mayoría de los casos, que alrededor del 10 % al 15 % de los miembros de la comunidad tenga un nivel de participación activa alto, entre un 15 % al 20 % sean miembros moderadamente activos, mientras que el resto posea un nivel bajo o nulo de participación [2]. En el presente trabajo hemos medido los niveles de participación a través del uso de técnicas de análisis de redes sociales, utilizando los *logs* de actividad y la base de datos del sistema de software utilizado por la comunidad, y los comparamos con los niveles de participación esperados.

- Identidad [46][97]: Los miembros de una comunidad, a través del tiempo van generando un lazo "invisible" que los une, y que hace que sientan que pertenecen a una estructura social en común. Este sentido de identidad es lo que empuja a los miembros de una comunidad dada a participar en forma voluntaria, y es el elemento que hace que la comunidad sea una estructura viva y que en si misma tenga una identidad global propia. En el presente trabajo se propone medir la identidad percibida por los miembros a través de encuestas a los miembros de la comunidad.

- **Propósito** [5][98][96]: Una comunidad virtual existe por un motivo, por algún propósito que motiva a sus miembros a formar parte de esa comunidad inicialmente. Sin embargo, sus miembros podrían desarrollar propósitos diferentes. Las razones por la cuales una persona pertenece a una comunidad varían. Algunos quieren información o soporte, interactuar con otros, entretenerse, conocer nuevas personas, o escuchar sus propias ideas. Cada participante tiene su propio motivo. Entendiendo los aspectos que motivan a las personas a entrar y retornar a una comunidad, es posible tomar decisiones técnicas y sociales adecuadas. La

comunidad debe generar un balance entre estos propósitos, de tal forma que cada miembro pueda cumplir sus propios objetivos, manteniendo el objetivo común que define a la comunidad. Para un mejor entendimiento, hemos dividido esta dimensión en los siguientes indicadores:

- Cumplimiento: El fenómeno más esperable dentro de una comunidad es que todos sus miembros busquen la realización de un objetivo común [2]. De esta forma, todas las interacciones y esfuerzos existentes son para desarrollar una o varias ideas comunes. En el presente trabajo se propone medir el cumplimiento utilizando encuestas a los administradores (para definir los objetivos de la comunidad), en conjunto con técnicas de minería de texto, para analizar lo que efectivamente dicen las personas en la comunidad, y si existe relación con los objetivos definidos por los administradores.
- Claridad: Si bien una comunidad puede tener definido un objetivo común que mueve a sus miembros a trabajar en forma conjunta, sus miembros pueden percibir este objetivo común de diferentes formas. Mientras para algunos puede ser claro, para otros puede no serlo, especialmente para los miembros nuevos [5]. En el presente trabajo se propone medir la claridad a través de encuestas a los miembros de la comunidad (para medir la claridad percibida por dichos miembros).

- **Políticas** [96][46][97]: Uno de los aspectos fundamentales dentro de una comunidad son las labores de administración. Las comunidades necesitan políticas para dirigir y sistematizar el comportamiento de sus miembros. Específicamente, las políticas son necesarias para establecer los requerimientos para ingresar a una comunidad, definir el estilo de comunicación entre los participantes, las conductas aceptadas, y otras. Los moderadores deben guiar a la comunidad para que pue-

da desarrollarse a través del tiempo, a través de tareas específicas que contribuyan a que la comunidad alcance sus objetivos. Sin embargo, el nivel de políticas existente puede afectar directamente la forma en la cual participan los miembros de una comunidad, al romper el balance con los otros aspectos sociales, por ejemplo, al utilizar políticas muy restrictivas o permisivas. Específicamente, en el presente trabajo de Tesis se medirá el siguiente indicador:

1. Nivel de Moderación: Se refiere a los esfuerzos hechos por los moderadores necesarios para mantener su funcionamiento. En el presente trabajo se propone medir los esfuerzos de moderación a través del análisis de los *logs* de actividad y de la base de datos del sistema de software utilizado por la comunidad.

Un hecho fundamental es que estas dimensiones sociales, se interrelacionan. Esto provoca que, en la búsqueda del desarrollo en una esas dimensiones, se esté afectando inevitablemente a las otras dos. Por ejemplo, en una comunidad virtual, que trate sobre cierto temas específicos, como la confección de artículos electrónicos. En la medida que se realizan acciones para aumentar la permisividad en los contenidos de los mensajes intercambiados entre usuarios, el propósito efectivo de la comunidad y de sus miembros podría divergir rápidamente.

La situación anterior se refleja en la Figura 2.5. Las 3 dimensiones sociales se relacionan entre sí, ocasionado que los cambios provocados en cualquiera de ellas, pueda tener un efecto en las otras dos, los cuales, podrían ser positivos o negativos, dependiendo de la naturaleza del cambio, y el tipo de comunidad sobre el cual se efectúa.

Las interrelaciones existentes entre las dimensiones sociales señaladas, resaltan la necesidad de incluir todos estos aspectos en cualquier estrategia de evaluación relacionada con la infraestructura computacional a comunidades virtuales. Sólo de esa forma será posible comprender la naturaleza del

Figura 2.5: Interrelación de aspectos sociales que posee los sistemas de apoyo a comunidades virtuales.

comportamiento de las comunidades. La omisión del estudio de algunos de estas dimensiones, podría omitir información importante relacionada con la comunidad virtual, no solamente información ligada a dicha dimensión en particular, sino además información sobre las relaciones e impactos con las otras dimensiones.

Es importante destacar que la influencia de una dimensión sobre otra, puede generar un efecto positivo o negativo que depende, entre otras cosas, de la naturaleza de la comunidad. Por ejemplo:

- Si el moderador de la comunidad espera aumentar los niveles de participación de los miembros de la comunidad, a través de un mecanismo de libre participación, en el cual cualquiera puede escribir lo que quiera en el foro de discusión que utiliza la comunidad. Si el moderador omite el hecho de la existencia del propósito, podría suceder que efectivamente en un comienzo exista mayores niveles participación, pero absolutamente fuera del propósito de la comunidad, ocasionado ideas

divergentes. Dicha situación puede provocar a largo plazo el efecto contrario al buscado: que las personas dejen de participar, puesto que la comunidad ya no les ofrece un espacio para lograr sus objetivos, dado que el propósito de la comunidad cambia con el tiempo, de acuerdo a cómo se comporten sus miembros.

- Si el moderador de la comunidad quiere aumentar los niveles de moderación de la comunidad, con el objetivo de que converja el propósito, es decir, se hablen mayoritariamente de los temas relacionados con el objetivo común de la comunidad, y no de otros, el efecto inmediato que podría provocar es la disminución en la participación. Sin embargo, a largo plazo, dado el propósito especifico de la comunidad, muchas más personas se animarían a participar si efectivamente las comunicaciones están orientadas al objetivo común.

Lo anterior muestra que es importante conocer estas interrelaciones, puesto que determinarán comportamientos que quizás no sean necesariamente los esperados en una comunidad, y que por tanto, es necesario considerarlos en una etapa temprana de la creación de la red social.

Capítulo 3

Casos de estudio

Este capítulo describe los casos de estudio utilizados como demostradores del trabajo experimental de esta Tesis. Estos casos de estudio nos han proporcionado colecciones de datos de su funcionamiento que nos permiten realizar los experimentos computacionales reportados en el Capítulo 5.

3.1. Comunidad Virtual I

Esta Comunidad Virtual de Práctica está formada por un grupo de personas que se han reunido con el objetivo de compartir conocimiento y experiencias en la construcción de efectos de música, amplificadores y equipos de audio. Los miembros se caracterizan por tener un enfoque hacia "hazlo tú mismo" (DIY – Do It Yourself). En un comienzo los miembros de la comunidad se enfocaban principalmente a la construcción y utilización de plexies (un amplificador clásico con el cual es posible obtener un sonido diferente al que se obtiene con amplificadores tradicionales). Sin embargo, se desarrollaron nuevos temas de interés con el transcurso de los años, tales como la lutería (la confección de instrumentos musicales de cuerda, como por ejemplo, una guitarra), audio profesional, compra/venta de partes, etc.

La comunidad virtual cuenta con aproximadamente 14 años de existencia,

en los cuales han participado más de 2.500 miembros. La interacción entre los miembros de la comunidad es realiza a través de un sistema de foro de discusión. En la actualidad se encuentra en una etapa en la cual la participación de sus miembros ha ido decayendo considerablemente en los últimos años. Al comienzo de la existencia de la comunidad la tarea de administración era una tarea más bien sencilla y era efectuada por un solo miembro. En el período de mayor actividad de la comunidad virtual dicha tarea tuvo que ser realizada por varios administradores.

3.1.1. Análisis de la Comunidad

Se dispone de los datos de comunidad, desde sus comienzos el año 2002, hasta mediados del año 2015. Los miembros de esta comunidad disponen básicamente de dos formas de interactuar con otros miembros (a) a través de la publicación de mensajes en uno de los foros de discusión de la comunidad, o (b) mediante del envío de mensajes privados a otros miembros de la comunidad. Tal como se visualiza en la Figura 3.1, la mayor cantidad de interacciones se da entre los años 2007 y 2008. El estudio de esta comunidad se encuentra enfocado en los mensajes que se publican en los diferentes foros de la comunidad, los cuales corresponden a casi la totalidad de las interacciones entre los miembros de la misma.

En la Tabla 3.1 es posible apreciar cómo se distribuyen las participaciones de los miembros de la comunidad en los diferentes foros de discusión, ya sea (a) creando un nuevo hilo de discusión (llamados topics), o (b) escribiendo un mensaje en un hilo de discusión existente (llamados posts).

La mayoría de los foros fueron creados en los inicios de la existencia de la comunidad virtual, con excepción de los foros de “Audio Pro”, “Sintetizadores” y “Hardcore DIY”, los cuales fueron creados en los últimos años, siendo este último foro de acceso restringido sólo para ciertos miembros de la comunidad.

Otro aspecto importante que es posible visualizar en esta comunidad vir-

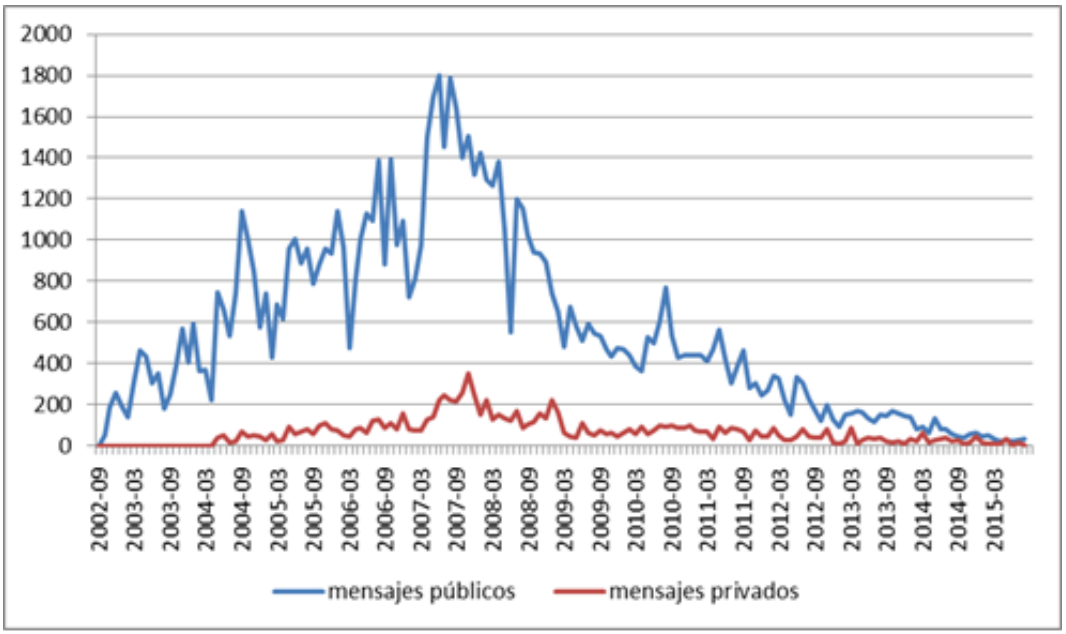

Figura 3.1: Actividad de la Comunidad I.

Cuadro 3.1: Participación de los miembros en los foros de discusión de la Comunidad I

Foro	# posts	# topics	%
Amplificadores	22.918	2.705	26,3 %
Efectos	32.062	3.470	36,8 %
Luthería	10.083	1.399	11,6 %
General	19.886	2.707	22,8 %
Audio Pro	1.850	201	2,1 %
Sintetizadores	327	35	0,4 %
Hardcore DIY	117	12	0,1 %
Total	87.243	10.529	100 %

Cuadro 3.2: Estadísticas de los hilos de discusión de la Comunidad I

Año	# hilos de discusión	# palabras	promedio
2002	84	6.384	76
2003	471	18.184	38,61
2004	916	26.696	29,14
2005	1.487	32.513	21,86
2006	1.690	36.601	21,66
2007	1.883	49.974	26,54
2008	1.415	41.270	29,17
2009	969	27.271	28,64
2010	876	29.285	33,43
2011	646	25.317	39,19
2012	412	18.368	44,58
2013	277	13.137	47,43
2014	197	7.321	37,16
2015	59	3.089	52,36
Total	11.382	335.890	29,51

tual, es que la participación en el foro "General" (creado para discusiones generales, no necesariamente relacionadas con el objetivo de la comunidad virtual) alcanza el 22,8 % del total de interacciones, mientras que la participación en los otros foros (relacionados con temas específicos de la comunidad virtual) corresponde al 77,2 %. La Tabla 3.2 da la estadísticas de los hilos de discusión.

En la Figura 3.2 se muestra la cantidad de miembros que participan activamente en la comunidad a lo largo de los diferentes años de existencia de la comunidad virtual. Es posible observar que la mayor cantidad de miembros activos se encuentran entre los años 2006 y 2009.

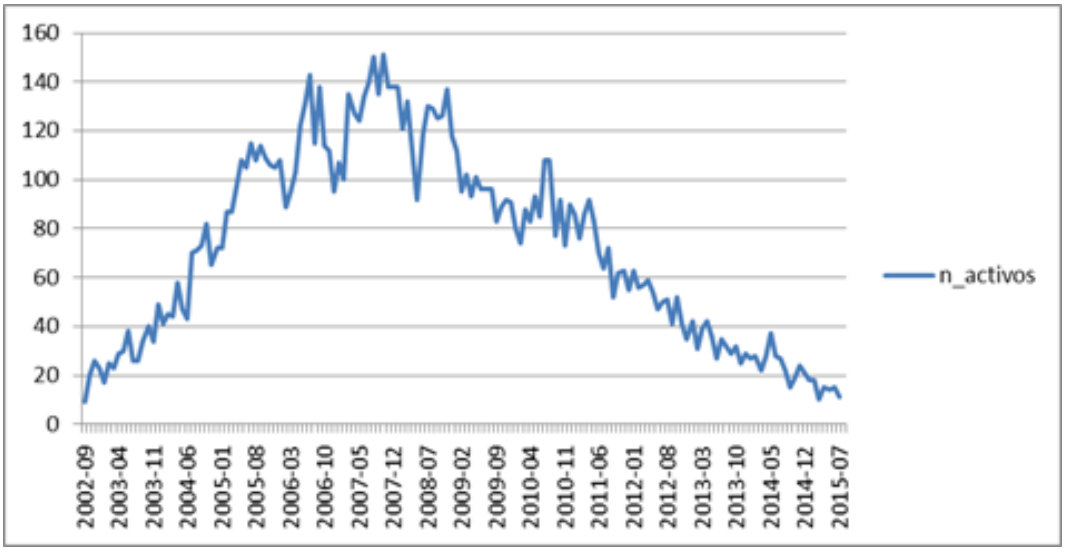

Figura 3.2: Total de miembros activos de la Comunidad I, basados en el total de interacciones.

3.2. Comunidad Virtual II

Corresponde a una comunidad virtual de práctica creada en el interior de una empresa, con el objetivo de crear e intercambiar ideas de innovación. Sus miembros pueden escribir sus ideas y el resto de los miembros puede opinar al respecto. Todas las interacciones entre los miembros de esta comunidad son realizadas a través del uso de una infraestructura de comunicaciones creada especialmente para ellos.

La comunidad cuenta con más de 500 miembros en sus más de tres años de experiencia. La tarea de administración es realizada por un empleado externo contratado especialmente para dicha actividad.

3.2.1. Análisis de la Comunidad

Se dispone de los datos de la comunidad, desde su formación el año 2013, hasta comienzos del año 2015. La forma de interacción entre los miembros es

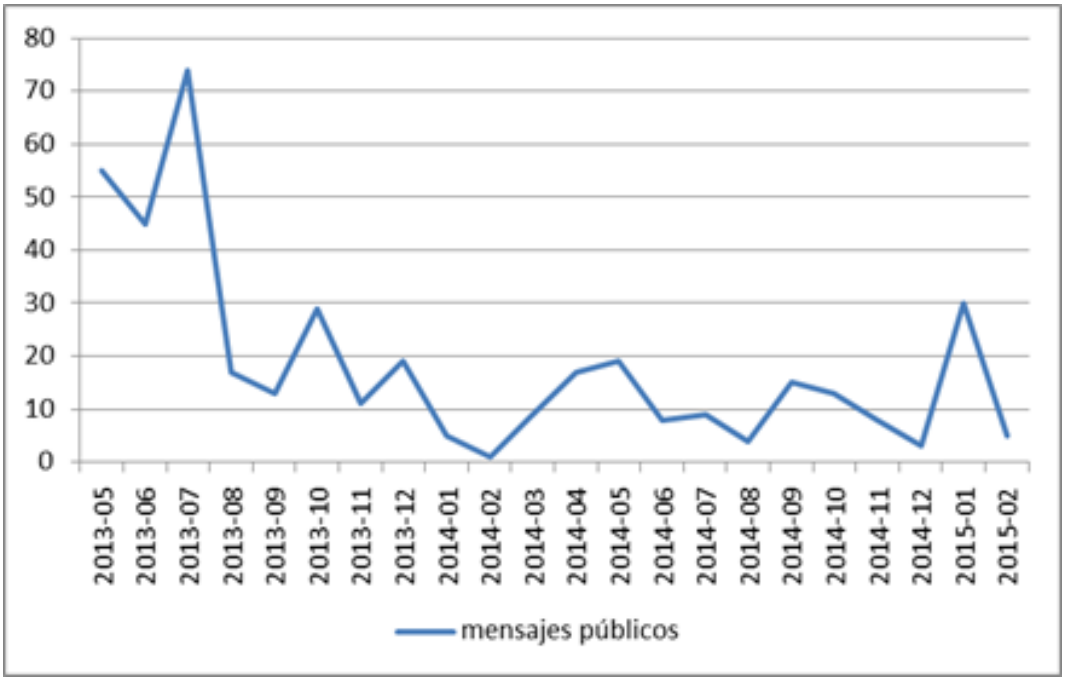

Figura 3.3: Actividad de la Comunidad II.

a través de la publicación de comentarios a ideas escritas por otros miembros. En la Figura 3.3 es posible visualizar el total de interacciones entre los años 2013 y 2015.

En la Figura 3.4 se muestran los miembros de la comunidad que participan activamente. Es posible observar que la mayor participación se concentra al comienzo de la existencia de la comunidad, y que posteriormente la participación permanece baja durante el resto de los años.

Al igual que la comunidad I, se generan los archivos necesarios para el procesamiento. En la Tabla 3.3 se resumen estadísticas sobre hilos de discusión extraídas de dichos archivos.

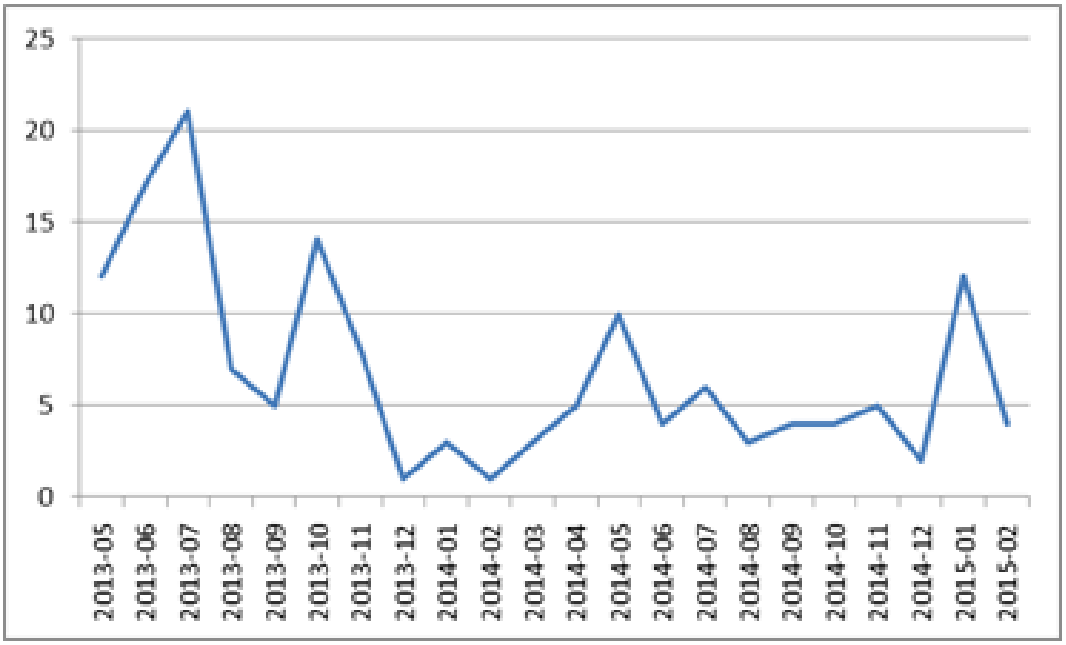

Figura 3.4: Total de miembros activos de la Comunidad II, basados en el total de interacciones.

Cuadro 3.3: Estadísticas de los hilos de discusión de la Comunidad II

Año	# hilos de discusión	# palabras	promedio
2013	207	2.725	11,49
2014	77	1.682	21,84
2015	26	707	27,19
Total	11.382	335.890	29,51

Capítulo 4

Aproximaciones al modelado estructural/semántico

En este capítulo se presentan los métodos computacionales de os que nos hemos servido para el análisis de una Red Social en Línea (OSN, por sus siglas en inglés), incorporando la explicación de los aspectos metodológicos más relevantes. Primero, se muestra una descripción general del objetivo del modelado de la red social, para posteriormente explicar con detalle cada paso y método computacional que se considera. Introducimos los conceptos básicos del analisis semántico de documentos de texto y las dos aproximaciones que hemos probado, una basada en lógica difusa y la otra en modelos gráficos probabilísticos. A continuación presentamos las técnicas seguidas para generar la red social a partir de las comunicaciones entre los miembros, seguido del proceso de filtrado de la red. Por ultimo referimos la construcción del grafo de la red social guiado por la información semántica siguiendo tres aproximaciones distintas.

4.1. Descripción General

Una OSN se define por las relaciones sociales que se establecen entre sus miembros. Estas relaciones son la base sobre la cual se crea el sentimiento compartido de pertenencia, y sobre la cual se crea una identidad común entre sus miembros. Para que puedan mantenerse estas relaciones sociales es necesario que sus miembros interactúen en forma sostenida a través del tiempo, de tal forma que se consoliden los lazos que los unen y no sean sólo personas aisladas que forman un grupo sin identidad.

Típicamente, para poder evaluar las relaciones sociales que se forman se utilizan algoritmos de análisis de redes sociales (SNA por sus siglas en inglés). En la medida en que la participación de una persona en la OSN produce interacciones y respuestas de otros miembros, se considera mucho más valiosa que la de aquellos miembros que participan en forma aislada. Los algoritmos de SNA pueden medir ese fenómeno a través del uso de métricas de centralidad, determinando cuán importante es una persona respecto de la red en la que participa, pudiendo identificar diferentes niveles de participación dentro de la OSN. De esta forma, un miembro posee una importancia mayor, y por tanto se le puede considerar como un miembro clave dentro de la OSN, si genera más relaciones que otros miembros.

Esto es aún más relevante para algunos tipos específicos de OSN, como lo son las Comunidades Virtuales de Práctica (VCoP), en las que también es importante que las interacciones entre sus miembros cumplan un determinado propósito. Sin ello, la comunidad sólo sería un grupo de personas interactuando sin ningún objetivo. Dicho propósito es lo que los une, y por lo tanto, pertenecen a ella porque comparten un interés común. Esto último genera que los miembros también desarrollen un lenguaje común, que tiene directa relación con la existencia de la comunidad.

El objetivo de la metodología de modelado propuesta es mejorar el descubrimiento de los niveles de participación de los miembros de una OSN, específicamente la detección de los miembros claves. Para ello, se propone la

realización conjunta del análisis estructural de la de red social subyacente y el análisis semántico de los contenidos generados por sus miembros. esto es, queremos mejorar las mediciones que se realizan sobre el comportamiento de la comunidad virtual, al combinar en el análisis las relaciones que forman las "Personas", y el hecho de que tengan un "Propósito" al interactuar entre si (ver Figura 4.1).

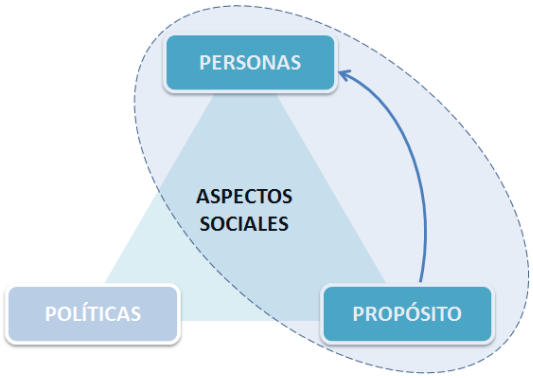

Figura 4.1: Objetivo del modelo propuesto, desde la perspectiva de los aspectos sociales definidos en [Aguilera2017].

El enfoque propuesto se basa en obtener una representación reducida / filtrada de la red social interna. Sin embargo, esta representación debe crearse de tal manera que la información contenida sea mejor para descubrir a los miembros clave (alineados con los objetivos de las redes sociales y los miembros que producen interacción). El siguiente paso es aplicar un algoritmo de centralidad sobre los miembros de comunidad, como el algoritmo HITS [Kleinberg1999] por ejemplo, para obtener los miembros clave de la OSN en función de sus respectivos temas de interés. Como resultado, se obtiene el rango de los miembros completos de OSN donde, para cada tema interesante de la red social, es posible revelar miembros que pueden ser considerados como expertos o clave en esos temas.

En las OSN ocurre frecuentemente que un número reducido de miembros son responsables de generar gran parte de los flujos de información de la

red social completa [101]. Este es un hecho conocido en el mundo de las comunidades virtuales. Wenger [39] plantea que aproximadamente del 10 al 15 % de los miembros de una Comunidad de Práctica pertenecen al núcleo de la comunidad, y otro 10 al 15 % pertenece a los miembros activos. Lo anterior señala que entre un 20 al 30 % de la comunidad son los miembros más importantes de la comunidad, siendo el resto (del 70 al 80 %) parte de la periferia de los miembros, es decir, corresponden a miembros que sólo participan en forma ocasional o son completamente pasivos.

Tal como se mencionó anteriormente, el primer paso para la identificación de los miembros clave de la red es agregar semántica a la estructura de la red. Para realizar esto, hemos considerado dos estrategias, una basada en el uso de conceptos (CB) [Rios2008], y otra basada en la identificación de tópicos mediante el uso de dos técnicas, *Latent Dirichlet Analysis* (LDA) y PYTM. Mientras que CB es un proceso semiautomático, LDA y PYTM son procesos automáticos para la detección de tópicos en documentos de texto.

Las OSN usualmente son soportadas por sistemas que permiten la interacción entre sus miembros, como por ejemplo a través del uso de foros de discusión (como por ejemplo VBuletin, PHPbb, etc.). En el enfoque SNA tradicional, se utilizan los mensajes del foro para crear una estructura de grafo (dirigido o no dirigido). En dicha representación, es posible determinar que los miembros de la red corresponden a los nodos y las relaciones entre dichos nodos representa cuando un miembro le responde a otro. De esta forma, si un miembro A responde un mensaje publicado por un miembro B, se considera como una interacción entre los nodos A y B. En un OSN, muchas veces, los usuarios simplemente usan el sistema para escribir publicaciones que están lejos de los propósitos u objetivos principales de OSN, por lo tanto, estas publicaciones pueden incluso molestar a otros miembros.

Es por esto que se propone que al dibujar el grafo de la estructura de la red definida por las respuestas, es conveniente realizar previamente un análisis semántico de su contenido. Si están alineados con los propósitos principales

de OSN, entonces es una interacción positiva y debe mantenerse, en otro caso, puede descartarse como una fuente de ruido. Como resultado, una vez que se aplican las técnicas basadas en conceptos o tópicos para agregar semántica a las publicaciones, la red resultante es una versión filtrada de la red original. Para hacerlo, se puede filtrar por un tema específico y eliminar interacciones que pertenecen a otros temas, o podemos establecer un umbral para filtrar información irrelevante, etc.

4.2. Procesamiento del texto

A continuación se introducen algunos conceptos que son utilizados más adelante. Primero, se define $\mathcal{V}$ como un vector de palabras que contiene el vocabulario a ser utilizado. Se define una palabra w, como una unidad básica de dato discreto, indexado por $\{1, ..., |\mathcal{V}|\}$. Un mensaje publicado es una secuencia de S palabras definida por $\mathbf{w} = (w^1, ..., w^S)$, en donde w^s representa la s^{th} palabra dentro del mensaje. Finalmente, un corpus es definido como un conjunto de $\mathcal{P}$ mensajes denotado como $\mathcal{C} = (\mathbf{w}_1, ..., \mathbf{w}_{|\mathcal{P}|})$.

Una representación vectorial de los mensajes del corpus está dada por `TF-IDF`$= (m_{ij}), i \in \{1, \ldots, |\mathcal{V}|\}$ y $j \in \{1, \ldots, |\mathcal{P}|\}$, en donde m_{ij} es el peso asociado y representa si una palabra dada es más importante que otra dentro de un documento. Los pesos m_{ij} considerados en este trabajo son una mejora sobre los definido en [Salton1975] como *tf-idf* (*term frequency times inverse document frequency*), y se definen como sigue:

$$m_{ij} = \frac{n_{ij}}{\sum_{k=1}^{|\mathcal{V}|} n_{kj}} \times log\left(\frac{|\mathcal{C}|}{n_i}\right), \tag{4.1}$$

donde n_{ij} es la frecuencia de la i^{th} palabra en el j^{th} documento, y n_i es el número de documentos que contiene la palabra i. El término *tf-idf* es la representación, basada en el peso, de la importancia que tiene una palabra en un documento que pertenece a una colección de documentos. El

término *tf* (*term frequency*) indica el peso de cada palabra en un documento, mientras que el término *idf* (*inverse document frequency*) indica si la palabra es frecuente o poco común en el documento, estableciendo un peso alto o bajo respectivamente.

4.2.1. Lógica Difusa para Clasificación basada en Conceptos

Con el objetivo de clasificar el texto siguiendo una estrategia basada en concepto, se utilizó la metodología propuesta en [Rios2006] que hace uso de variables lingüísticas.

Los valores de las Variables Lingüísticas (LV) no son números sino palabras u oraciones en lenguaje natural. Sea u una LV, es posible obtener un conjunto de términos $T(u)$ los cuales cubren su universo de discurso U. Por ej. $T(temperatura) = \{frio, agradable, calor\}$ o $T(presion) = \{alta, normal, baja\}$.

Con el objetivo de utilizar LV para clasificación basada en conceptos, se asume que un documento puede ser representado como una relación difusa $[Concepts \times WP]$, $[C \times WP]$ en notación compacta, la cual es una matriz en donde cada fila es un concepto y cada columna es un mensaje (Web Post). Para obtener dicha matriz se reescribe esta relación de una forma mas conveniente en la Ecuación 4.2.

$$[C \times WP] = [Concepts \times Terms] \otimes [Terms \times WP], \quad (4.2)$$

donde "Terms" denomina a las palabras que pueden ser utilizadas para definir un concepto y "WP" se refiere a cualquier palabra dentro de un mensaje. En la Ecuación (4.2) los simbolos "$\times$" y "$\otimes$" representan la relación difusa y la composición difusa respectivamente.

Recordamos que $|\mathcal{P}|$ denota el número total de mensajes en el sitio completo, $|\mathcal{V}|$ el número total de palabras diferentes pertenecientes a todos los documentos, y K el número de conceptos definidos por el sitio utilizado por

la OSN. Es posible caracterizar la matriz $[C \times WP]$ a través de su función de pertenencia difusa especificada en la Ecuación 4.3

$$\mu_{C\times WP}(x,z) = \begin{pmatrix} \mu_{1,1} & \mu_{1,2} & \cdots & \mu_{1,|\mathcal{P}|} \\ \mu_{2,1} & \mu_{2,2} & \cdots & \mu_{2,|\mathcal{P}|} \\ \vdots & \vdots & \vdots & \vdots \\ \mu_{K,1} & \mu_{K,2} & \cdots & \mu_{K,|\mathcal{P}|} \end{pmatrix}, \tag{4.3}$$

en donde $\mu_{C\times WP} = \mu_{[C\times T]\otimes[T\times WP]}$ representa la función de pertenencia de la descomposición difusa en la Ecuacion 4.2, y cuyos valores están en el rango $[0, 1]$.

Existen diversas alternativas para realizar la composición difusa. En [Nakanishi1993] se presenta un estudio entre seis diferentes modelos. Un aspecto importante que debe ser considerado es que, incluso si algunos términos no están presentes en un mensaje, el grado de expresión de un concepto no debe sufrir alteraciones. Este es el motivo para utilizar la regla de composición dada por la Equación 4.4 [Nakanishi1993].

$$\mu_{Q\circ Z} = \bigvee\{\mu_Q(x,r) \wedge \mu_Z(r,y)\} \tag{4.4}$$

Sean $Q(U,V)$ y $Z(V,W)$ dos relaciones difusas que comparten un conjunto común V. Sea $\mu_Q(x,r)$ con $x \in U \wedge r \in V$ y $\mu_z(r,y)$ con $r \in V \wedge y \in W$ funciones miembro para Q y Z respectivamente, entonces se puede escribir la regla de composición como se muestra en la Ecuación 4.4, en donde $\bigvee$ es la suma limitada definida por $min(1, x + r)$ y $\wedge$ es el producto algebraico $= (x \cdot r)$.

Para aplicar el método definido anteriormente, es necesario identificar los conceptos relevantes para el estudio. Es importante remarcar que no se busca una clasificación conceptual como la utilizada en el área de recuperación de información, la cual podría requerir la inclusión de miles de conceptos y términos para recuperar todos los documentos relevantes independientemente de las palabras clave utilizadas en la consulta del usuario. Lo que se busca

son conceptos que describan el cumplimiento de los miembros al propósito de la red social. Para identificar cuáles son los conceptos más interesantes para describir el comportamiento de los miembros en la OSN se utiliza el conocimiento de los expertos de la red social.

Posteriormente, con el uso de diccionarios se extraen términos para definir los conceptos relevantes, es decir, para expresar cada concepto como una lista de términos (asumiendo que un concepto es un LV). Se utilizan sinónimos, cuasi-sinónimos, antónimos, etc.

Después, se define las funciones de pertenencia para las relaciones difusas $[Concepts \times Terms]$ y $[Terms \times WP]$. Se utiliza la frecuencia relativa de las palabras en un mensaje para representar los valores de pertenencia difusa de la matriz $[Terms \times WP]$.

Un poco más compleja es la construcción de los valores $[Concepts \times Terms]$. Se realiza pidiendo al experto que asigne el grado en que cada término representa a un concepto. Para ello, el debe comparar dos términos cada vez y dar un valor entre 0 y 1. Por ejemplo, un sinónimo puede recibir un valor cercano a 1; un cuasi-sinónimo podría recibir un valor entre 0.65 y 1; un antónimo puede ser definido como 0, etc. Finalmente, se obtiene se obtiene la relación difusa $\mu_{G \times P}(x, z)$ aplicando la Ecuación 4.4.

4.2.2. Proceso de Extracción de Tópicos usando modelos gráficos

Un modelo de tópico puede considerarse como un modelo probabilístico que relaciona documentos y palabras a través de variables que representan los tópicos principales inferidos del texto mismo. En este contexto, un documento se puede considerar como una mezcla de tópicos, representados por distribuciones de probabilidad que pueden generar las palabras en un documento con estos tópicos. El proceso de inferencia de las variables latentes, o tópicos, es el componente clave de este modelo, cuyo objetivo principal es aprender la distribución de los tópicos subyacentes en un corpus dado de

documentos de texto.

Latent Dirichlet Allocation (LDA)

Uno de los modelos de tópicos mas utilizados es el Latent Dirichlet Allocation (LDA) [Blei2003]. Corresponde a un modelo Bayesiano en el cual se infieren los tópicos latentes de los documentos desde las distribuciones de probabilidad estimadas a partir del conjunto de datos de entrenamiento. La idea principal de LDA, es que cada tópico es modelado por una distribución de probabilidades sobre un conjunto de palabras representadas por un vocabulario ($w \in \mathcal{V}$), y cada documento es modelado por una distribución de probabilidades sobre un conjunto de tópicos ($\mathcal{T}$). Asumimos que los documentos siguen distribuciones multinomiales de Dirichlet.

La ventaja de este método sobre el enfoque basado en conceptos, es que es un proceso automático, sólo se necesita a los expertos para proveer un nombre a cada tópico descubierto por el algoritmo.

Dados los parámetros de suavizado β y α, y una distribución conjunta de una mezcla de temas θ, la idea de LDA es determinar la distribución de probabilidades de generar, a partir de un conjunto de tópicos $\mathcal{T}$, un mensaje compuesto por un conjunto de S palabras denotado por $\mathbf{w} = (w^1, ..., w^S)$,

$$p(\theta, z, \mathbf{w}|\alpha, \beta) = p(\theta|\alpha) \prod_{s=1}^{S} p(z_s|\theta) p(w^s|z_s, \beta) \tag{4.5}$$

donde $p(z_s|\theta)$ puede ser representado por la variable aleatoria θ_i, tal que el tópico z_s se encuentra presente en el documento i ($z_s^i = 1$). Una expresión final puede ser deducida integrando la Ecuación 4.5 sobre la variable aleatoria θ y sumando sobre los tópicos $z \in \mathcal{T}$. Dado esto, la distribución marginal de un mensaje puede ser definida como sigue:

$$p(\mathbf{w}|\alpha, \beta) = \int p(\theta|\alpha) \left(\prod_{s=1}^{S} \sum_{z_s \in \mathcal{T}} p(z_s|\theta) p(w^s|z_s, \beta) \right) d\theta \tag{4.6}$$

El objetivo final de LDA es estimar las distribuciones descritas anteriormente para construir un modelo generativo para un corpus dado de mensajes. Existen diversos métodos desarrollados para efectuar la inferencia sobre estas distribución de probabilidades, tales como *variational expectation-maximizatio*n [Blei2003], que corresponde a una aproximación variacional discreta de la ecuación 4.6 empíricamente utilizada en [**?**], y por un modelo Gibbs sampling Markov chain Monte Carlo implementado y aplicado eficientemente en [xxxxxxx].

Pitman-Yor Topic Model (PYTM)

Es un modelo basado en LDA, el cual usa un proceso Pitman-Yor [xxxxxx] para generar la distribución *a priori*. Dicho proceso es una distribución sobre un espacio de probabilidades que usa tres parámetros: un parámetro de concentración, un parámetro de descuento d $(0 < d < 1)$ y una distribución base G_0. El proceso Pitman-Yor es una generalización del proceso Dirichlet, en el cual el parámetro de descuento es considerado como cero [xxxxx]. Este modelo se basa en el *Chinese Restaurant Process* (CPR), correspondiente a un proceso en donde n clientes se sientan en un restaurante chino con un número infinito de mesas. Siempre el primer cliente se sienta en la primera mesa. Los siguientes clientes se sientan en una mesa ocupada o en una nueva mesa desocupada. Para PYTM la representación del CRP está compuesta por 4 elementos: un cliente, una mesa, un plato y el restaurante. El cliente es representado por una palabra en un documento. La mesa es representada por una variable latente. El plato es representado por un tipo de palabra. El restaurante es representado por un documento. Dado que si se aumenta el valor de n aumenta la cantidad de mesas ocupadas, muchas con un sólo cliente, el proceso CRP resulta útil para capturar la distribución de ley de potencia (*power law*) que siguen las palabras.

La representación del modelo es mostrada en la Figura 4.2.

En el modelo, se define que:

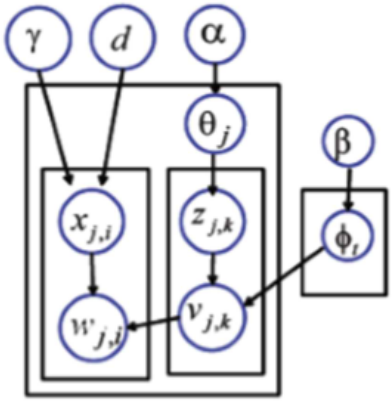

Figura 4.2: Representación del modelo de tópicos Pitman-Yor (Figura obtenida de [114]).

- $x_{j,i} = k$, indica que el i-simo cliente (palabra) se encuentra sentado en la k-sima mesa del restaurante (documento) j.
- $v_{j,k} = v$, indica que el plato (tipo de palabra) v es servido en la k-sima mesa del restaurante (documento) j.
- $w_{j,i}$ corresponde al i-sima cliente (palabra) del restaurante (documento) j.
- $z_{j,k}$ corresponde al tópico asignado a la k-sima mesa en el restaurante (documento) j.

4.3. Generación de la Red Social

Para construir la red social, antes de ser filtrada, se debe tener en cuenta la interacción de los miembros. En general, la actividad de los miembros se sigue de acuerdo con su participación en el foro. La Figura 4.3 ilustra la dinámica del sistema de respuestas en el foro. De esta forma, la participación de un miembro se manifiesta cuando un publica un mensaje en la red social. Debido

a que la actividad de la OSN se describe de acuerdo con la participación de sus miembros, la red se configura de la siguiente forma: los nodos corresponden a los miembros del OSN y los arcos representan la interacción entre ellos.

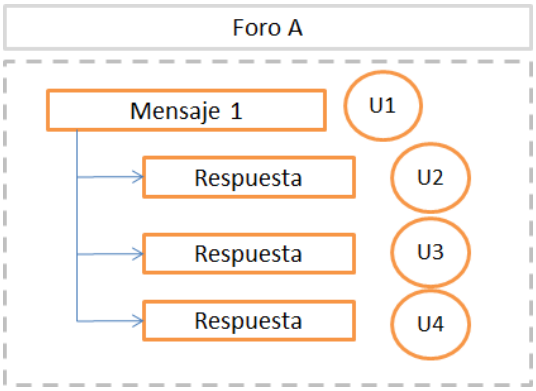

Figura 4.3: Una estructura típica de un foro de discusión, en donde los círculos corresponden a los usuarios que publican mensajes.

En este trabajo, se categoriza la OSN de acuerdo al significado que tienen la respuestas de los miembros:

1. **Red Orientada al Creador (creator)**: Cuando un usuario A escribe un mensaje en una discusión X, dicho mensaje es una respuesta dirigida sólo al miembro que inició la discusión. Esta representación de red es la menos densa (en donde la densidad es medida en términos del número de arcos que tiene la red).

2. **Red Orientada a la Última Respuesta (reply_prev)**: Cuando un usuario A escribe un mensaje en una discusión X, dicho mensaje es una respuesta dirigida sólo al último mensaje publicado independientemente de su originador. Esta forma de representación posee una densidad media.

3. **Red Orientada a Todas las Respuestas Previas (reply_all)**: Cuando un usuario A escribe un mensaje en una discusión X, dicho

mensaje es una respuesta a todos los mensajes anteriores. Esta forma de representar la red es la más densa.

4. **Red Orientada a Últimas Respuestas Previas (reply_all30)**: Cuando un usuario A escribe un mensaje en una discusión X, dicho mensaje es una respuesta sólo a los mensajes anteriores que hayan sido escritos con una antigüedad de a lo más 30 días.

En la figura 4.4 se despliegan los tres enfoques de representación de red social del foro de discusión. Los arcos representan respuestas de los miembros y los nodos representan a los miembros quienes realizaron un mensaje. Utilizando un enfoque tradicional, el peso de cada arco es la cuenta simple de cuántas veces un miembro dado responde a otro.

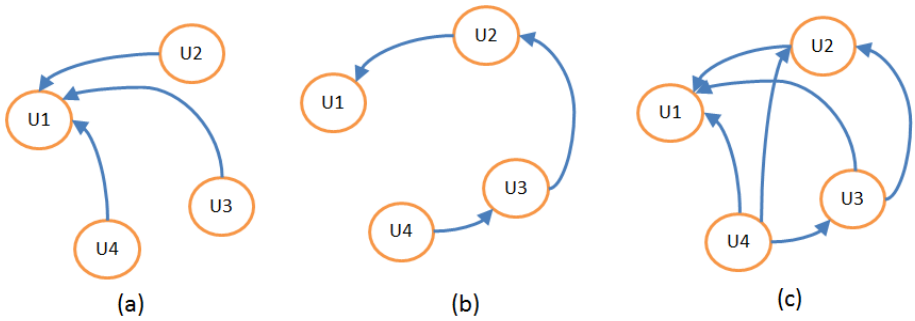

Figura 4.4: Tres diferentes modelos de representación en forma de red las interacciones dentro de un foro. (a) creator, (b) reply_prev, (c) reply_all

Con el objetivo de considerar solamente aquellas respuestas de miembros que se encuentran relacionadas con el propósito de la red social (para cualquiera de las configuraciones presentadas), y para además filtrar los mensajes que corresponden a ruido", se aplica a continuación reducción de mensajes en base al uso de técnicas basadas en análisis de tópicos y conceptos.

4.4. Filtrado de la Red Social basado en Conceptos & Tópicos

Siguiendo la aproximación presentada en [Rios2009], el cual provee un mecanismo para evaluar el cumplimiento a los objetivos de la comunidad, se crea un proceso para clasificar los mensajes de los miembros de una comunidad de acuerdo a los objetivos de la OSN. Estos objetivos son definidos como un conjunto de términos compuestos por un conjunto de palabras clave u oraciones en lenguaje natural. Dichos términos son definidos mediante la técnica basada en conceptos o las técnicas basadas en tópicos.

La idea principal de esta aproximación es reducir la densidad de la representación en grafo, considerando solo los mensajes que son relevantes al propósito de la red social, y los hilos de conversación que se relacionan entre si. De esta forma, si un miembro realiza una pregunta, se buscan todos los mensajes que respondan esa pregunta siguiendo el del hilo de discusión. Para lo anterior, se calcula cuán similares son las respuestas respecto del mensaje al cual están respondiendo. Para ello, se utiliza la función de similitud coseno. Si esta similitud tiene un valor sobre cierto umbral θ prefijado, se considera que existe una interacción entre ambos. Posteriormente, se reduce la densidad de la red manteniendo los vectores de los mensajes que son similares al tópico/concepto componente que ellos tienen. Posteriormente, se suman todas las interacciones para los mensajes entre un miembro A y un miembro B para cada hilo de discusión dentro del foro.

$$ARC\left(U_A, U_B\right) \;=\; \sum_{\tau=1}^{T} d_\tau\left(\mathbf{P}_A\;,\;\mathbf{P}_B\right) \tag{4.7}$$

En la Ecuación 4.7, se define la función ARC, que calcula el peso de la interacción entre el usuario A y el usuario B. En dicha ecuación, $\mathbf{P}_A$ son todas las interacciones (o mensajes) de un miembro A en un hilo de discusión y $\mathbf{P}_B$ es lo mismo para el miembro B. Ambas, son las interacciones siguiendo

un hilo específico τ el cual está en el rango $1 \ldots T$, siendo T el último hilo dentro del foro.

$$d_\tau\left(\mathbf{P}_A\ ,\ \mathbf{P}_B\right)\ =\ \sum_{i=1}^{|P_A|}\sum_{j=1}^{|P_B|} d\left(\mathbf{P}_{A\tau i}\ ,\ \mathbf{P}_{B\tau j}\right)\quad d\left(\mathbf{P}_{A\tau i}\ ,\ \mathbf{P}_{B\tau j}\right)\ \geq\ \theta\quad \forall \mathbf{P}_{A\tau i},\ \mathbf{P}_{B\tau j} \tag{4.8}$$

$$d(\mathbf{P}_i, \mathbf{P}_j) = \frac{\sum_k w_{ik} w_{jk}}{\sqrt{\sum_k w_{ik}^2 \sum_k w_{jk}^2}} \tag{4.9}$$

donde $w_{\tau ik}$ es el puntaje de un tópico/concepto k en un mensaje del usuario i definido por la *Semantic Weights Matrix* (`SWM`) (ver Algoritmo 1) en un hilo de discusión τ. Se calcula $d_m(P_{\tau i}, P_{\tau j})$ sólo si P_j es una respuesta a P_i. Después de eso, los pesos del arco $a_{i,j}$ son calculados de acuerdo a si la distancia es más grande que el umbral θ, agregando un $+1$ al hilo de discusión τ por la interacción entre el nodo i y el nodo j.

Finalmente, para evitar aquellos mensajes que poseen pocas palabras, o mensajes generales como "si", "chao", "me ayuda mucho", se eliminan. Para ello, en el cálculo de la función $ARC(.)$ no se consideran los mensajes cuyos componentes no son relevantes en al menos un cierto nivel ρ, en donde $\rho \in (0, 0{,}2]$.

El enfoque anterior se aplica a las cuatro configuraciones de red social presentadas anteriormente (creator, reply_prev, reply_all y reply_all30). Posteriormente se aplican diferentes algoritmos de centralidad, específicamente PageRank, HITS, in-degree, out-degree, con el objetivo de encontrar a los miembros claves en las diferentes configuraciones de redes.

4.5. Algoritmo de Centralidad

A continuación se enumeran las técnicas de centralidad que son utilizadas:

Degree Centrality

El grado de un vértice $d(v)$ es una medida que cuenta el número de aristas incidentes a v.

$$d(v) = \sum_{j}(a_{ij}) \tag{4.10}$$

Mediante la evaluación de la centralidad de esta forma, se compara la conectividad de los vértices sin medir ni mostrar qué tan bien posicionado se encuentran los vértices dentro del grafo.

Closeness Centrality

Es una medida que evalúa cuan cercano se encuentra un vértice a los otros vértices de un grafo. Si $dist(s, v)$ es la distancia más corta entre los vértices s y v, la centralidad de un vértice v se calcula de la siguiente forma:

$$c(j) = \sum_{s \in G \setminus \{v\}} \frac{1}{dist(s, v)} \tag{4.11}$$

Betweenness Centrality

La idea principal de esta métrica es que los vértices que se encuentran en el camino geodésico (aquel camino de longitud menor entre todos los posibles caminos que unen dos nodos en un grafo) de muchos otros vértices poseerán un gran control sobre el flujo de información debido a que ellos se encuentran "entre" (*between*) los otros vértices. La centralidad de un vértice v para esta métrica se calcula como:

$$b(j) = \sum_{s \in V \setminus \{v\}} \sum_{t \in V \setminus \{s,v\}} \frac{\sigma_{st(v)}}{\sigma_{st}} \tag{4.12}$$

En donde σ_{st} es el número de caminos geodésicos entre los vértices s y t. El valor $\sigma_{st(v)}$ corresponde al número de caminos geodésicos entre s y t, en

los cuales v pertenece a dichos caminos.

Hubs y *Authority*

Son dos indicadores que son obtenidos a partir del algoritmo HITS, el cual es utilizado para valorar, y de paso clasificar, la importancia de una página web. *Authority* valora cuan buena es una página como recurso de información. *Hub* dice cuan buena es la información que se consigue siguiendo los enlaces que tiene a otras páginas. Ambas métricas pueden ser utilizadas para calcular la centralidad de vértices al interior de un grafo.

Pagerank

También corresponde a un algoritmo diseñado originalmente para medir la importancia de páginas web. La centralidad $P(j)$ es una variante de la centralidad basada en vectores propios y puede ser determinada por la siguiente ecuación:

$$P(j) = \frac{(1-d)}{n} + d \sum_{i \in B_j} \frac{P(i)}{|F_i|} \tag{4.13}$$

En donde d es probabilidad de saltar en forma ocasional de una página a otra, n corresponde al total de nodos (páginas en este caso), F_i el conjunto de vértices (páginas) a los cuales el vértice i apunta, B_i el conjunto de vértices (páginas) que apuntan a i. Dado que es una función recursiva, se evalúa de forma iterativa hasta que el valor de $P(j)$ converja.

4.6. Construcción del grafo pesado de Red Social guiado por información semántica

La base del procedimiento se presenta en el Algoritmo 1 que muestra cómo determinar la *Semantic Weights Matrix*, la cual asigna un puntaje a cada mensaje de acuerdo a todos los conceptos o tópicos considerados para la construcción de la red social. El algoritmo inicialmente determina la matriz `TF-IDF` de acuerdo a la Ecuación 4.1, la matriz semántica `SM` de acuerdo a los tópicos o conceptos basados en el proceso de procesamiento de texto de la sección 4.2.2 y la sección 4.2.1, respectivamente. Finalmente, la multiplicacion de las matrices entre `TF-IDF` y `SM` define la matriz `SWM`.

Algorithm 1 Initialize Semantic Weights Matrix

Require: $\mathcal{V}$ (Vocabulary)
Require: $\mathcal{P}$ (Posts)
Require: k (Number of Topics or Concepts)
Ensure: Semantic Weights Matrix $\texttt{SWM}[|\mathcal{P}|, k]$
1: $\texttt{TF-IDF}[|\mathcal{P}|, |\mathcal{V}|]$ (Eq. 4.1)
2: $\texttt{SM}[k\,\mathcal{V}] \leftarrow$ Build `SM` (semantic matrix) according to `Topics` (Sec. 4.2.2) or `Concepts` (Sec. 4.2.1)
3: $\texttt{SWM}[|\mathcal{P}|, k] \leftarrow \texttt{TF-IDF} \times \texttt{SM}^{\mathrm{T}}$

El Algoritmo 2 presenta el pseudo-código de creación del grafo $\mathcal{G}_c = (\mathcal{N}, \mathcal{A})$ el cual es construido usando el enfoque de red oriento al creador del mensaje. Primero, se construye la matriz SWM de acuerdo al Algoritmo 1. A partir de ella, tomando en cuenta todos los mensajes $\mathcal{P}$, se construye la red siguiendo la estructura presentada en Figura 4.4 (a). Esto es, para cada mensaje i, el peso del arco $a_{i,j}$ es incrementado de acuerdo al número de respuestas j, si la distancia entre los mensajes es más grande o igual al umbral θ.

El Algoritmo 3, al igual que el anterior, construye en primer lugar la

matriz `SWM`. Entonces, siguiendo la estructura de la Figura 4.4 (b), para cada mensaje de un usuario i, el arco $a_{i,j}$ se incrementa de acuerdo al número de todas las respuestas j cuya distancia entre mensajes sea mayor o igual al umbral θ.

Finalmente, el Algoritmo 4, define la matriz `SWM` y entonces construye la red *All-Previous-oriented*. De acuerdo a la Figura 4.4 (c), para cada mensaje de un usuario i, el peso del arco $a_{i,j}$ es incrementado de acuerdo al total de respuestas directas j cuyos distancia entre mensajes sea mayor o igual al umbral θ.

Algorithm 2 Creator-oriented Network

Require: $\mathcal{P}$ (Posts)
Ensure: Network $\mathcal{G}_c = (\mathcal{N}, \mathcal{A})$
1: Build `SWM` according to Algorithm 1
2: Initialize $\mathcal{N} = \{\}, \mathcal{A} = \{\}$
3: **for** each $i \in \mathcal{P}$ **do**
4: $\quad \mathcal{N} \leftarrow \mathcal{N} \cup i$
5: **end for**
6: **for** each $i \in \mathcal{P}.creator$ **do**
7: $\quad$ **for** each $j \in \{i.replies\}, i \neq j$ **do**
8: $\quad\quad$ **if** $d_m(P_i, P_j) \geq \theta$ **then**
9: $\quad\quad\quad a_{i,j} \leftarrow a_{i,j} + 1$
10: $\quad\quad\quad \mathcal{A} \leftarrow \mathcal{A} \cup a_{i,j}$
11: $\quad\quad$ **end if**
12: $\quad$ **end for**
13: **end for**

Algorithm 3 Reply-oriented Network

Require: $\{\mathcal{V}, \mathcal{P}, k\}$
Ensure: Network $\mathcal{G}_c = (\mathcal{N}, \mathcal{A})$
1: ... (as presented in algorithm 2)
2: **for** each $i \in \mathcal{P}$ **do**
3: **for** each $j \in \{i.reply\}, i \neq j$ **do**
4: **if** $d_m(P_i, P_j) \geq \theta$ **then**
5: $a_{i,j} \leftarrow a_{i,j} + 1$
6: $\mathcal{A} \leftarrow \mathcal{A} \cup a_{i,j}$
7: **end if**
8: **end for**
9: **end for**

Algorithm 4 All-Previous-oriented Network

Require: $\{\mathcal{V}, \mathcal{P}, k\}$
Ensure: Network $\mathcal{G}_c = (\mathcal{N}, \mathcal{A})$
1: ... (as presented in algorithm 2)
2: **for** each $i \in \mathcal{P}$ **do**
3: **for** each $j.ReplyTo(i), j \in \mathcal{P}, i \neq j$ **do**
4: **if** $d_m(P_i, P_j) \geq \theta$ **then**
5: $a_{i,j} \leftarrow a_{i,j} + 1$
6: $\mathcal{A} \leftarrow \mathcal{A} \cup a_{i,j}$
7: **end if**
8: **end for**
9: **end for**

Capítulo 5

Experimentos, Resultados y Evaluación

Este capítulo presenta los resultados de los experimentos computacionales realizados sobre datos de redes sociales reales como demostradores de las técnicas propuestas en la presente Tesis. Primero, introducimos las métricas utilizadas para la evaluación comparativa y la metodología de evaluación de los resultados obtenidos. Posteriormente, se presentan los resultados obtenidos al aplicar los modelos presentados en el Capítulo 4. Presentamos resultados para grafos de las dos redes sociales descritas en el Capítulo 3 construidos sin realizar filtro semántico (USN: Unfiltered Social Network), filtradas en base a conceptos difusos (FCA: Fuzzy Concept Analysis), filtradas en base a tópicos latentes (LDA), y filtradas en base a tópicos de un proceso PYTM.

5.1. Métricas y Metodología de Evaluación

Con el objetivo de evaluar la calidad el modelo propuesto, se definió el siguiente marco metodológico de evaluación (Ver Figura 5.1).

Primero, se calculan las densidades de las redes sociales subyacentes, an-

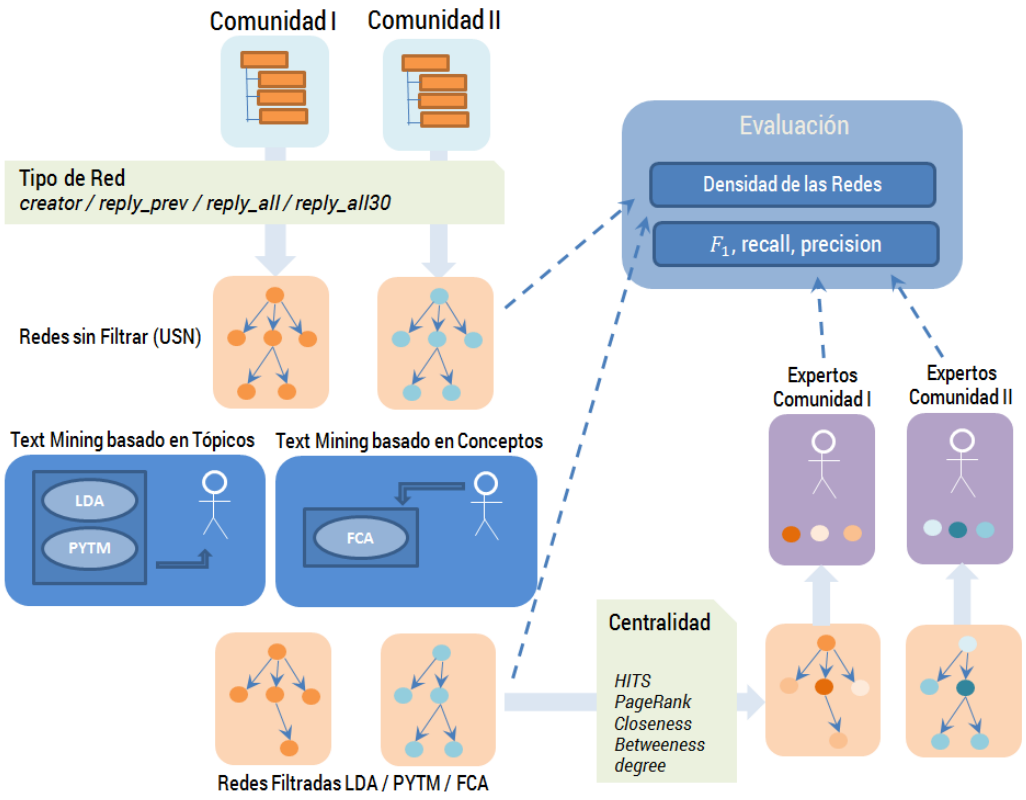

Figura 5.1: Marco metodológico de evaluación.

tes de aplicar el modelo, con el objetivo de medir cuánto se reduce la densidad de las 4 formas de representación de red (*creator*, *reply_prev*, *reply_all* y *reply_all30*) al filtrar semánticamente mediante el uso conceptos y tópicos.

$$densidad(G) = \frac{|E|}{|V|(|V|-1)} \tag{5.1}$$

Posteriormente, se aplica el modelo propuesto agregando semántica a las 4 formas de representación de red social, utilizando las tres técnicas presentadas en el Capítulo 4 (FCA, LDA y PYTM).

Los algoritmos de LDA y PYTM se calibran utilizando la métrica de perplejidad, la cual corresponde a una medida para comparar distribuciones

de probabilidad. Esta medida captura qué tan bien una muestra es predicha por un modelo. Basada en la entropía de la distribución subyacente, esta medida identifica cuán difícil es para el modelo escoger una palabra en la distribución. Denotamos D como la colección de documentos, M como el número de documentos, N_d como el número de palabras en el documento d, y $p(w_d)$ como la probabilidad asignada por el modelo a una palabra w_n, esta medida se calcula según la ecuación siguiente:

$$perplexity(D) = exp\{-\frac{\sum_{d=1}^{M} log\ p(w_d)}{\sum_{d=1}^{M} N_d}\} \tag{5.2}$$

Un valor menor de perplejidad indica un rendimiento general mejor del modelo. Con esta métrica se establece la cantidad de tópicos para las técnicas PYTM y LDA.

Para evaluar la calidad de los resultados obtenidos con cada combinación experimental, se le solicita a el(los) administrador(es) de la red social que identifiquen, para un período determinado, un listado de los miembros que ellos consideran como centrales. Sólo se le(s) solicita a el(los) administrador(es) dicha información para un período determinado, entre 1 a 2 años, dado que son humanos y no recuerdan necesariamente la historia completa de la red social.

Dado que para el(los) administrador(es) no es posible ordenar por importancia el listado de miembros, se les solicita que identifiquen 4 grupos de la siguiente forma:

- Expertos Tipo A: corresponden a los miembros más importantes de la comunidad.
- Expertos Tipo B: corresponden a miembros menos importantes que los del Tipo A, sin embargo siguen siendo miembros clave.
- Expertos Tipo C: corresponden a miembros que fueron importantes en el pasado, y que generalmente participaron desde el comienzo de la red

social, sin embargo en la actualidad no participan mucho.

- Expertos Tipo X: corresponden a miembros que no son importantes en la comunidad. Son miembros que no pertenecen al núcleo mas importante de la comunidad, y generalmente realizan más preguntas que respuestas.

Estos grupos son el *gold standard* contra el que validamos nuestros algoritmos. Aplicamos los diferentes algoritmos de centralidad para detectar a los miembros centrales de la estructura social en estudio. Se aplican a la red original, y a la red social filtrada por el contenido semántico. Posteriormente, se puede ordenar a los miembros de red social de acuerdo al ranking dado por el algoritmos de centralidad. Con ese listado, se seleccionan los primeros 10, 20, 30, etc. miembros más importantes y que se comparan con los miembros claves entregador por el(los) administrador(es). Para determinar el mejor rendimiento, se utiliza la métrica F_1 definida como:

$$F_1 = 2 * \frac{precision * recall}{precision + recall} \tag{5.3}$$

en donde

$$precision = \frac{|\{miembros\ claves\ definidos\ por\ administradores\} \cap \{nucleo\}|}{|\{nucleo\}|} \tag{5.4}$$

$$recall = \frac{|\{miembros\ claves\ definidos\ por\ administradores\} \cap \{nucleo\}|}{|\{miembros\ claves\ definidos\ por\ administradores\}|} \tag{5.5}$$

Cabe señalar que todos los resultados están basados en datos anónimos, para asegurar la privacidad de los miembros de las redes sociales en estudio.

5.2. Comunidad I

Se cuenta con los datos del foro de discusión utilizado por esta red social desde Septiembre del 2006 hasta Junio del 2014. Para realizar el estudio de densidad sobre esta comunidad, se utiliza completamente desde los años 2006 hasta el 2013. También se realiza el estudio para cada mes del año 2013. Para el caso de la evaluación de los miembros clave, se les solicita sólo para el año 2013 y 2014. Esta información fue solicitada el año 2014, cuando se realizó el primer estudio de esta comunidad, y se volvió a actualizar el año 2017. Dicha información fue provista por 2 de los 5 administradores existentes (los fundadores de la comunidad). Como resultado, se obtuvo una lista con 66 miembros identificados por los administradores, para los años 2013 y 2014, que fueron agrupados de la siguiente forma:

- Expertos Tipo A: Fueron identificados 34 miembros en este grupo para el año 2013.
- Expertos Tipo B: Fueron identificados 21 miembros en este grupo para el año 2013.
- Expertos Tipo C: Fueron identificados 11 miembros en este grupo para el año 2013.

5.2.1. Detección de tópicos y Definición de Conceptos

Dado que las técnicas de LDA y PYTM requieren como parámetro la cantidad de tópicos, las distintas configuraciones irán cambiando el total de tópicos que se debe calcular. Para ello, se calcula dichas técnicas desde 0 hasta 100 tópicos con incrementos de 5 (debido al alto tiempo de cálculo de estas técnicas). Adicionalmente, como se tiene sospecha de que la cantidad óptima de tópicos es menor que 25, se calculará también desde 1 a 25 tópicos, con incrementos de 1, de tal forma de que, cuando encontramos la cantidad

óptima de tópicos, está ya ha sido calculada sobre el conjunto de documentos. Dado que éstas técnicas tienen alto costo computacional, en especial la técnica de PYTM, es que a continuación, en la Figura 5.2, se plotean los tiempos de procesamiento de dicha técnica. Sólo esta técnica, requiere más de 372 horas de procesamiento. En la misma figura es posible visualizar que el tiempo de cálculo de la técnica PYTM crece más que la LDA conforme que se calcula una mayor cantidad de tópicos.

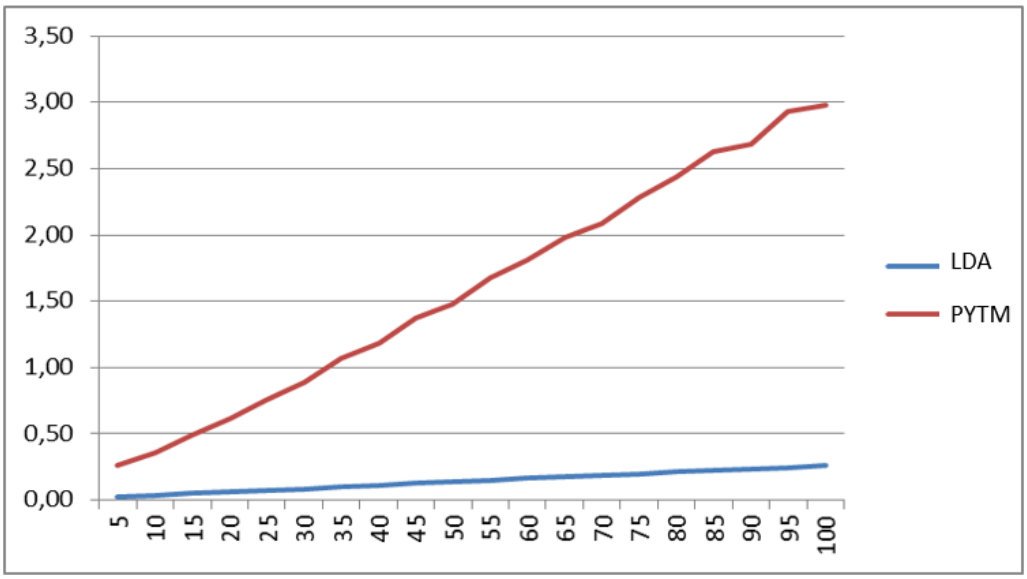

Figura 5.2: Comparación de tiempos de procesamiento entre LDA y PYTM, para el año 2013, con cantidad de tópicos entre 5 y 100, medido en horas.

Si bien la técnica de PYTM entrega mejores valores para los diferentes periodos en estudio, al aplicar la métrica *perplexity* (ver Figura 5.3), el filtrado semántico de las redes se realiza para ambas técnicas. Como fue mencionado anteriormente, con el resultado de la métrica *perplexity* se establece la cantidad de tópicos para cada período (ver Figura 5.4). En el Cuadro 5.1 se pueden ver dos tópicos de ejemplo, entregados por la técnica LDA para el año 2013.

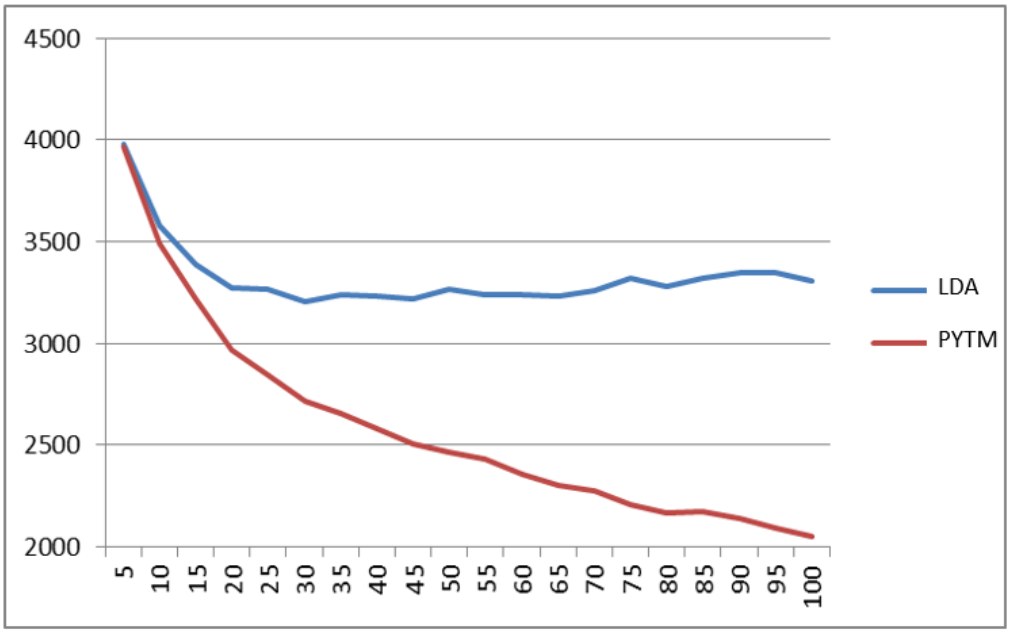

Figura 5.3: Comparación de técnicas LDA y PYTM en la Comunidad I, Año 2013, entre 5 y 100 tópicos.

Cuadro 5.1: Diez de las palabras más relevantes con sus respectivas probabilidades condicionales para dos tópicos obtenidos con LDA.

Tópico 5	Tópico 6
cable (0.02614838890648233)	placa (0.0185159180873805)
ruido (0.02518864483823386)	papel (0.012016382339933889)
cables (0.01703082025812189)	queda (0.009737324350569494)
tierra (0.014204907168279182)	caja (0.009568505240246206)
problema (0.0133518013298361)	placas (0.009315276574761272)
tiene (0.011485632308241858)	aluminio (0.007964723692174965)
guitarra (0.011219036733728394)	son (0.007373856806043454)
gracias (0.008872995678009919)	agua (0.007205037695720166)
volumen (0.008286485414080301)	después (0.006867399475073588)
jack (0.008233166299177608)	pintura (0.006867399475073588)

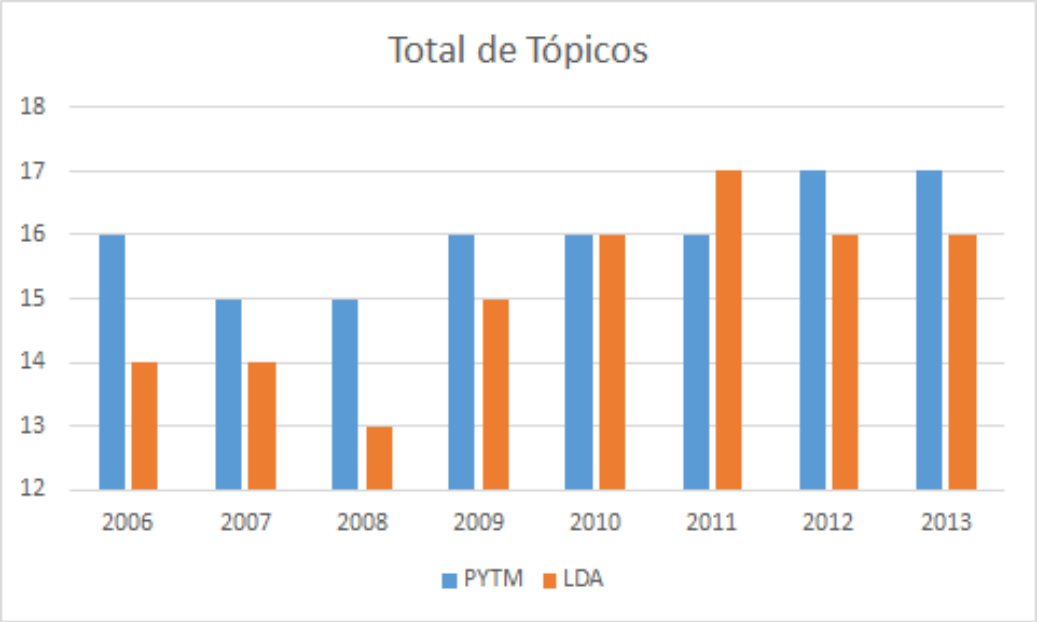

Figura 5.4: Total de tópicos para cada año de estudio para las técnicas PYTM y LDA.

5.2.2. Filtrado de arcos basado en conceptos y tópicos

El primer resultado interesante es que el filtro semántico reduce enormemente las densidades de la representación de grafo de la comunidad virtual, como se observa en la Fig. 5.5.

La información de todos los meses para todos los años entre el 2006 y 2013 fue utilizada para calcular la densidad de cada mes. Posteriormente, el resultado fue totalizado anualmente y desplegado en la Fig. 5.5. Es interesante observar que el método basado en concepto (FCA) siempre realiza una alta reducción de dimensión, independiente del tipo de red utilizada. Por otra parte, las redes filtradas por LDA y PYTM generan reducciones similares, generando la mayor reducción en las topologías orienta al creador y a todas las respuestas, lo cual puede verse en la Fig. 5.5.

Para mostrar aún mejor que los métodos basados en la semántica permi-

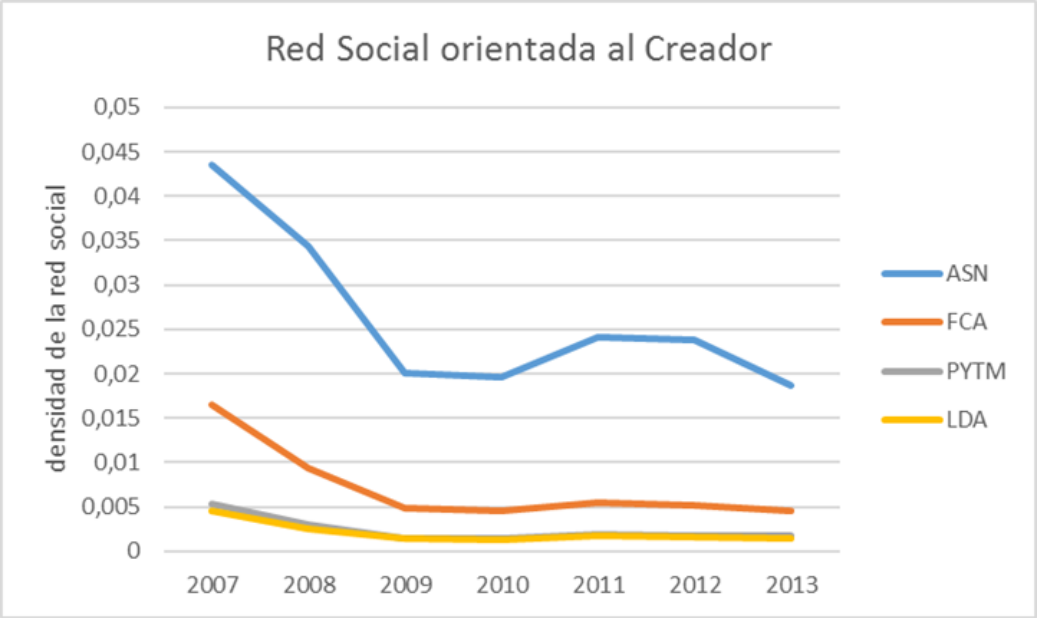

Figura 5.5: Reducción de la densidad de red desde los años 2006 al 2014 utilizando las 4 topologías de red.

ten reducir la densidad de una manera drástica, se despliega el grafo generado para todo el año 2013, para los tres métodos propuestos y el grafo original (sin aplicar ningún filtro) 5.6. De esta forma, es posible observar cómo las redes son filtradas por estos métodos. Adicionalmente, en Fig. 5.6, se aplican los métodos a las 4 formas de representar la interacción en un foro (*creator*, *reply_prev*, *reply_all* y *reply_all30*).

Se puede observar que en la primera fila de la Fig. 5.6 se encuentran las redes pertenecientes a la topología orientada al creador, y en la primera columna se encuentra la representación tradicional de la red (sin aplicar algún filtro semántico). En la segunda columna se encuentra la representación basada en conceptos utilizando un corte de 0,8 entre los mensajes, para mantener solo aquellos arcos que son realmente similares. De esta forma, si un mensaje que responde a otro tiene una similitud de 0,8, entonces ambos mensajes están relacionados y por lo tanto el arco se dentro de grafo se conserva. En

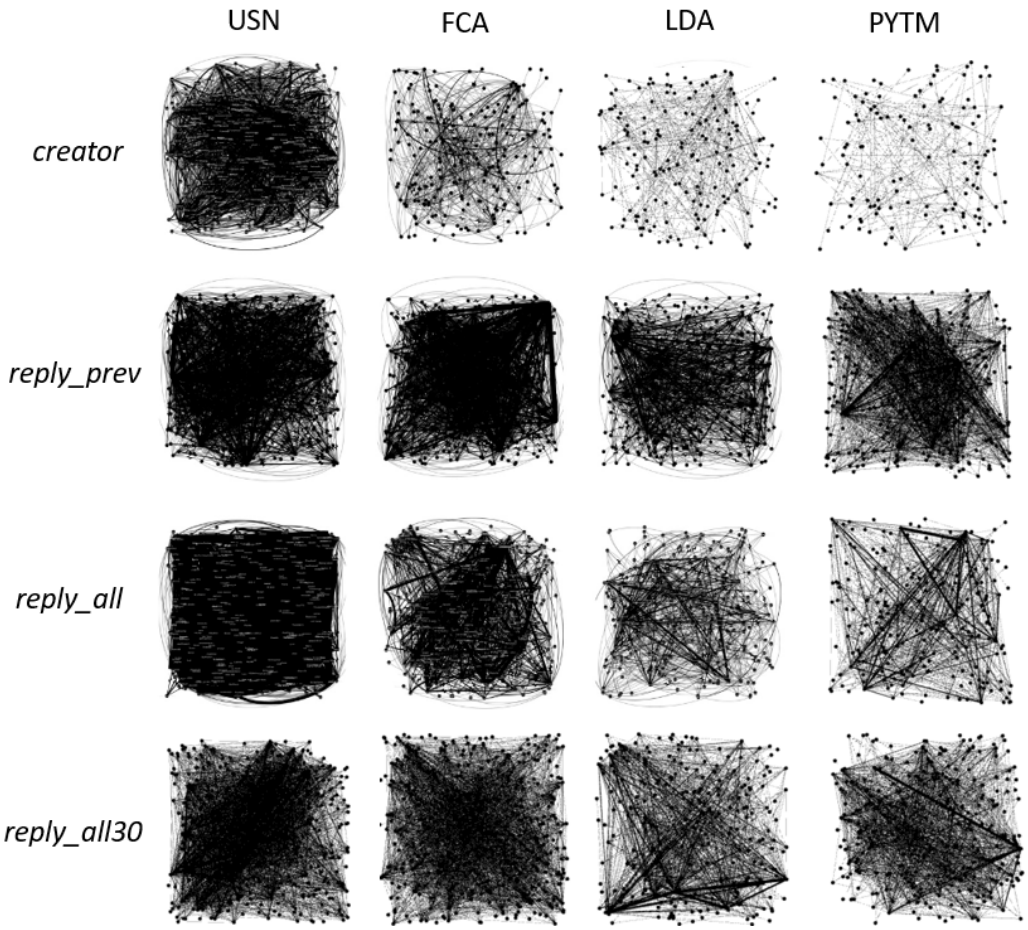

Figura 5.6: Reducción de red para los 3 métodos y todas las topologías para el año 2013.

caso contrario, el arco no es considerado, dado que el mensaje podría estar respondiendo a otro mensaje, o podría corresponder a un comentario poco útil que no se relaciona con el primer mensaje, y por lo tanto es mejor no considerarlo.

Finalmente, la mayor reducción de la representación de red es obtenida utilizando el modelo de tópicos LDA. En la Fig. 5.6 se utiliza el mismo corte que la para las redes basada en conceptos y la red basada en PYTM, el cual es 0,8 en ambos casos.

5.2.3. Detección de miembros claves

Como se mencionó anteriormente, los administradores clasificaron a los miembros claves en 3 grupos: A, B y C; se creó otro grupo, denominado X, de aquellos que no son miembros claves. De esta forma, denotamos como $A + B + C$ la detección de cualquier tipo de miembro clave. En la Fig. 5.7 se muestra una comparación entre la aplicación de HITS sobre las 4 formas de red (sin filtrar, en donde se representan todas las interacciones). Dado que HITS es un método de ranking, se seleccionaron los primeros 40 miembros con el ranking de hub más alto. De esta forma se compara con la lista entregada por los administradores. En dicha figura, se representan los miembros claves detectados. Se puede observar que la red *creator* es la que tiene mejor rendimiento. Lo anterior se repite para los otros algoritmos también, por lo cual en los siguientes análisis se muestran principalmente los cálculos aplicados sobre la red *creator*.

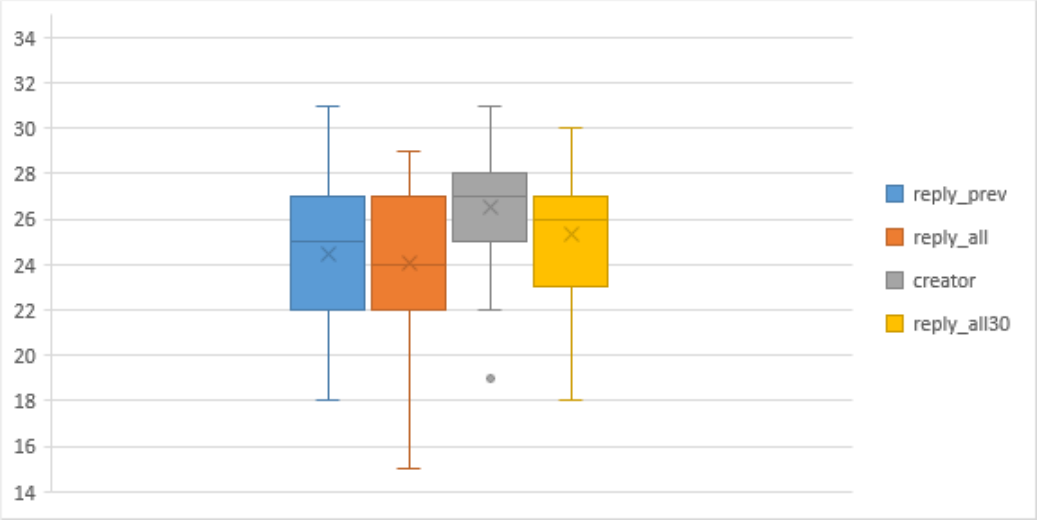

Figura 5.7: Comparación de las 4 redes Descubrimiento de miembros claves aplicados a los top 10, 20, 30, 40 miembros ordenaos por HITS hub.

En la Fig. 5.8 se muestra una comparación entre la aplicación de HITS sobre una red tradicional (sin filtrar, en donde se representan todas las inter-

acciones) y los resultados de HITS sobre la red filtrada utilizando los tópicos obtenidos con el método LDA. Se muestran los resultados seleccionando los primeros 10, 20, 30 y 40 miembros con el ranking de hub más alto. De esta forma se compara con la lista entregada por los administradores. En dicha figura, se representan los miembros claves detectados. La red utiliza corresponde a la red *creator*.

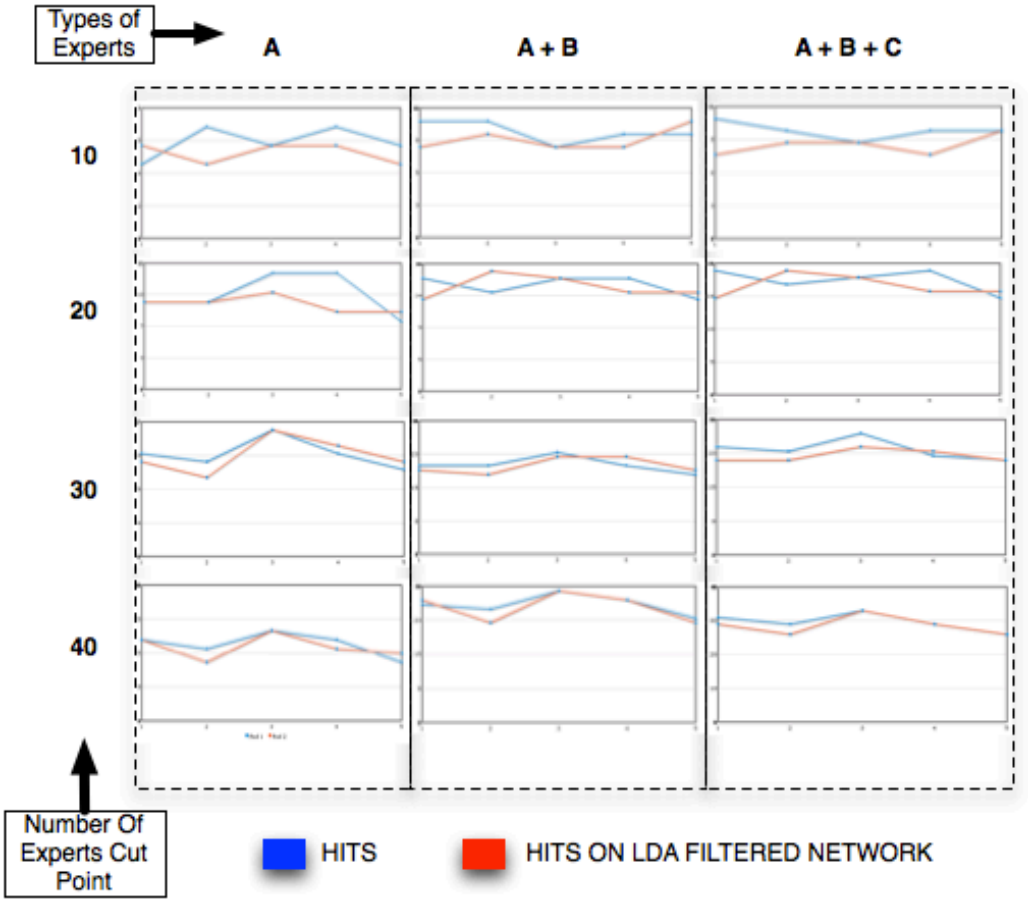

Figura 5.8: Descubrimiento de miembros claves aplicados a los top 10, 20, 30, 40 miembros ordenaos por HITS hub.

Los resultados muestran que las redes USN y LDA son similares en términos de expertos detectados. Sin embargo, no es posible afirmar que una de

ellas tiene siempre mejores resultados que la otra. La comparación en detalle de los últimos 5 meses del 2013 se puede ver en la Tabla 5.2. También es posible ver que las redes filtradas por FCA, LDA y PYTM tienen rendimiento similar. Sin embargo, si se mezclan los resultados obtenidos de HITS en la forma tradicional (USN) en conjunto con los resultados de la red basada en tópicos o conceptos, entonces el resultado en mejor que cualquier otro algoritmo. Lo anterior se puede ver en la tabla 5.2.

Cuadro 5.2: Detección de miembros claves para diferentes métodos y configuración de parámetros, para los últimos 5 meses del año 2013, utilizando la red *creator*

Tipo de Experto	Método	08-2013	09-2013	10-2013	11-2013	12-2013
A	HITS	18	16	20	18	13
	FCA 0.8/0.2	16	12	20	14	13
	HITS ∪ FCA 0.8/0.2	**22**	**19**	**23**	**18**	**16**
	PYTM 0.8	18	12	20	15	13
	HITS ∪ PYTM 0.8	**21**	**18**	**25**	**18**	**15**
	LDA 0.4/0.01	18	13	20	16	15
	HITS ∪ LDA 0.4/0.01	**21**	**19**	**25**	**19**	**16**
A+B	HITS	26	25	29	27	23
	CB 0.8/0.2	23	20	28	24	21
	HITS ∪ CB 0.8/0.2	**32**	**29**	**34**	**30**	**26**
	PYTM 0.8	26	21	29	26	20
	HITS ∪ PYTM 0.8	**30**	**29**	**36**	**30**	**25**
	LDA 0.4/0.01	27	22	29	27	22
	HITS ∪ LDA 0.4/0.01	**31**	**30**	**36**	**31**	**26**
A+B+C	HITS	31	29	33	29	26
	CB 0.8/0.2	27	25	34	26	23
	HITS ∪ CB 0.8/0.2	**38**	**34**	**40**	**33**	**29**
	PYTM 0.8	28	26	32	29	26
	HITS ∪ PYTM 0.8	**35**	**40**	**40**	**33**	**30**
	LDA 0.4/0.01	29	26	33	29	26
	HITS ∪ LDA 0.4/0.01	**36**	**40**	**41**	**33**	**30**

Un descubrimiento interesante es que los miembros detectados por HITS sobre la red sin filtrar (USN) son diferentes de aquellos descubiertos por HITS sobre la red filtrada por tópicos o conceptos. Por ejemplo, si se realiza la unión de los primeros 10 miembros detectados con la red USN con los primeros 10 miembros detectados por el método LDA, entonces se obtiene: (i) un conjunto de miembros que es descubierto por los dos métodos, (ii) un conjunto de miembros descubierto sólo por HITS sobre USN, y finalmente, (iii) una proporción de los miembros claves que sólo son descubiertos por la

red filtrada por LDA. Lo anterior significa que los métodos son complementarios, lo cual puede verse en la figura 5.9, en donde se puede ver la unión de los ranking que se obtienen desde HITS sobre la red sin filtrar y desde HITS filtrado por LDA. Se realizó esto para los primeros 10, 20, 30 y 40 miembros; también se realizó para descubrir por separado los miembros A, A+B y A+B+C. En la parte media de cada columna se encuentra la parte común de los resultados, es decir, miembros descubiertos por ambos métodos.

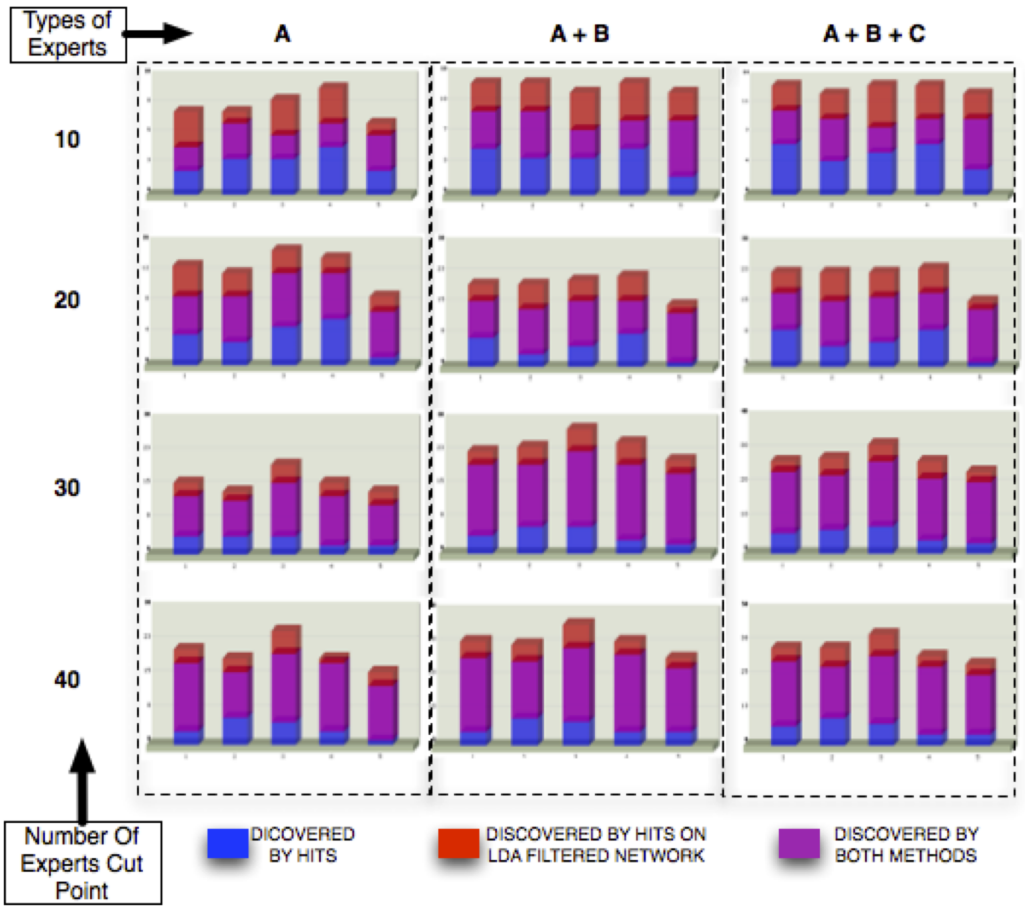

Figura 5.9: Métodos para descubrir miembros claves aplicados a los top 10, 20, 30, 40 miembros ordenados por HITS hub.

Es posible calcular *Precision*, *Recall* y el índice F para los rankings de

Cuadro 5.3: Rendimiento de la detección de miembros clave de Tipo A utilizando el método tradicional y la combinación con filtro LDA

Periodo	1	2	3	4	5
Precision	0.45	0.40	0.50	0.45	0.33
Precision HITS ∪ LDA	0.41	0.35	0.48	0.38	0.31
Recall	0.43	0.38	0.48	0.43	0.31
Recall HITS ∪ LDA	0.50	0.45	0.60	0.45	0.38
F	0.44	0.39	0.49	0.44	0.32
F HITS ∪ LDA	0.45	0.39	0.53	0.41	0.34

Cuadro 5.4: Rendimiento de la detección de miembros clave de Tipo $A + B$ utilizando el método tradicional y la combinación con filtro LDA

Periodo	1	2	3	4	5
Precision	0.65	0.63	0.73	0.68	0.58
Precision HITS ∪ LDA	0.61	0.55	0.69	0.62	0.50
Recall	0.47	0.45	0.53	0.49	0.42
Recall HITS ∪ LDA	0.56	0.55	0.65	0.56	0.47
F	0.55	0.53	0.61	0.57	0.48
F HITS ∪ LDA	0.58	0.55	0.67	0.59	0.49

Cuadro 5.5: Rendimiento de la detección de miembros clave de Tipo $A+B+C$ utilizando el método tradicional y la combinación con filtro LDA

Periodo	1	2	3	4	5
Precision	0.78	0.73	0.83	0.73	0.65
Precision HITS ∪ LDA	0.71	0.65	0.79	0.66	0.58
Recall	0.48	0.45	0.51	0.45	0.40
Recall HITS ∪ LDA	0.55	0.55	0.63	0.51	0.46
F	0.59	0.55	0.63	0.55	0.50
F HITS ∪ LDA	0.62	0.60	0.70	0.57	0.51

los miembros claves creados por diferentes métodos. En las tablas (5.3)(5.4) y (5.5) presentamos los rendimientos para el descubrimiento de miembros claves utilizando la red USN, y filtrado basado en tópicos obtenidos por LDA. Se utilizó un ajuste de $\theta = 0{,}4$ y $\rho = 0{,}01$. De las tablas mencionadas, se puede observar que los miembros claves descubiertos por la red sin filtrar tienen una *Precision* más alta que la combinación de USN y LDA. Sin embargo, las medidas de *Recall* y $F-measure$ son más alta para el algoritmo combinado. Esto es un buen resultado, debido a que es prioritario mejorar el *Recall*. En otras pruebas, cambiando el peso a los *ranking* resultantes para cada método, los resultados no cambian mucho, y son prácticamente los mismos.

Capítulo 6

Conclusiones y Trabajo Futuro

En este capítulo recogemos las conclusiones de esta Tesis y damos algunas indicaciones de trabajo futuro.

6.1. Conclusiones

En esta Tesis hemos estudiado un tipo específico de redes sociales en linea, las comunidades de práctica, que se caracterizan por tener un tema específico de interés para todos los miembros. Ese interés define un campo semántico muy concreto y acotado, por tanto es posible mediante técnicas de análisis semántico identificar las comunicaciones entre los miembros de la red social que son relevantes para este interés común y, por tanto, obtener una representación de la red social que se ajusta a la realidad de la actividad relevante en la comunidad. Usualmente, la red social inducida por una comunidad de práctica se construye en base a los mensajes publicados por los miembros. Esto es, en un hilo de discusión/comunicación consideramos que los mensajes definen las relaciones entre los miembros que participan en el hilo. Al menos hay tres formas en que se puede realizar esta construcción, y las hemos experimentado en esta Tesis. Hemos denominado filtrado de la red al proceso de análisis semántico de las comunicaciones en el sentido de que

reduce el número de conexiones efectivas de la red puesto que elimina conexiones debidas a mensajes irrelevantes. El objetivo de la construcción de la red es variado, pero hemos considerado dos utilidades: (a) la identificación de los miembros más importantes en la vida de la comunidad, esto es, aquellos que contribuyen de forma más abundante y creativa a los propósitos de la comunidad, y (b) el estudio de la evolución dinámica de la red y su estado de salud. Los trabajos en esta Tesis se han concentrado en el primero de estas dos utilidades.

Hemos definido dos técnicas de análisis semántico de los contenidos de los hilos de comunicaciones en la comunidad, una basada en lógica difusa que precisa la intervención de expertos, que denominamos basado en conceptos, y la segunda basada en la extracción de los tópicos latentes mediante algoritmos automatizados como son LDA y PYMT que fueron introducidos en el Capítulo 4. Como demostradores, hemos utilizado datos de dos comunidades de practica para las cuales además de los datos disponíamos de la información de los administradores relativa a los roles desarrollados por los miembros, i.e. la identificación de los miembros más relevantes desde el punto de vista del administrador del sistema. Esta información nos ha servido en ambas comunidades para contrastar y validar nuestras aproximaciones al filtrado de la red, obteniendo efectivamente representaciones filtradas cuyo análisis de centralidad obtiene la identificación de los miembros relevantes más aproximada a la realidad dada por los administradores del sistema

6.2. Trabajo futuro

Las aproximaciones de análisis semántico producen representaciones más fieles a la realidad del estado de las comunidades de práctica. Este tipo de análisis puede ser extendido al estudio de la evolución de la red, para detectar si se producen degeneraciones respecto del propósito inicial, permitiendo a los administradores tomar acciones correctivas. Las comunidades de prácti-

ca son limitadas por su propia naturaleza tanto en el numero de miembros como en los contenidos de los mensajes e intercambios de información. A pesar de esto, se trata de comunidades con un gran valor social cuya preservación es importante dentro y fuera de la comunidad, porque repercuten sus logros en el exterior de la comunidad. Por tanto, herramientas de control y mantenimiento como las que proponemos son doblemente valiosas.

Las técnicas de análisis semántico permiten, además, la indexación de los contenidos compartidos en la red, facilitando su recuperación. Esta minería del conocimiento proporciona un valor adicional a la comunidad y a sus miembros.

Por último, estas técnicas de análisis semántico/estructural pueden extenderse a otras redes sociales, como son las redes de recomendaciones que se crean en servicios web que ofrecen productos generales o productos muy específicos (e.g. películas, series). Los sistemas de recomendación actualmente se basan en medidas de similitud que no tienen en cuenta más que características objetivas de los productos, y no consideran el análisis semántico de las opiniones de los usuarios vertidas en comentarios o recomendaciones escritas en lenguaje natural. El uso de técnicas de inducción de tópicos mejoraría la creación de recomendaciones automáticas para alimentar la demanda.

Bibliografía

[1] P. Levy, Cyberculture. Minneapolis [u.a.]: Univ. of Minnesota Press, 2001.

[2] E. Wenger, R. McDermott, and W. Snyder, Cultivating Communities of Practice: A Guide to Managing Knowledge. Harvard Business School Press, 2002.

[3] A. J. Kim, Community Building on the Web: Secret Strategies for Successful Online Communities. 2000.

[4] F. Henri and B. Pudelko, "Understanding and analysing activity and learning in virtual communities," Journal of Computer Assisted Learning, vol. 19, pp. 474–487, Jan. 2003.

[5] J. Preece, "Online communities: Usability, sociabilty, theory and methods," 2001.

[6] J. Preece, "Etiquette, Empathy and Trust in Communities of Practice: Stepping-Stones to Social Capital," Journal of Universal Computer Science, Jan. 2004.

[7] B. Wellman, "An electronic group is virtually a social network," Culture of the Internet, Jan. 1997.

[8] E. Wenger, N. White, J. Smith, and K. Rowe, "Technology for communities," CEFRIO Book Chapter–Jan, Jan. 2005.

[9] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, and R. Wirth, "CRISP-DM 1.0 Step-by-step data mining guide," Aug. 2000.

[10] B. Wellman, J. Salaff, D. Dimitrova, L. Garton, M. Gulia, and C. Haythornthwaite, "Computer Networks as Social Networks: Collaborative Work, Telework, and Virtual Community," Annual Review of Sociology, vol. 22, no. 1, pp. 213–238, 1996.

[11] B. Wellman, "Computer Networks As Social Networks," Science, vol. 293, pp. 2031–2035, 2001.

[12] B. Wellman and M. Gulia, "Virtual communities as communities," Communities in cyberspace, Jan. 1999.

[13] C. Johnson, "A survey of current research on online communities of practice," The internet and higher education - Elsevier, Jan. 2001.

[14] J. Breslin and S. Decker, "The Future of Social Networks on the Internet: The Need for Semantics," IEEE Internet Computing, vol. 11, no. 6, pp. 86–90, 2007.

[15] B. Wellman, "Which Types of Ties and Networks Give What Kinds of Social Support?," Advances in Group Processes, vol. 9, pp. 207–235, 1992.

[16] R. Gross and A. Acquisti, "Information revelation and privacy in online social networks," Alexandria, VA, USA, 2005, pp. 71–80.

[17] danah michele boyd, "Friendster and publicly articulated social networking," Vienna, Austria, 2004, pp. 1279–1282.

[18] J. Donath and D. Boyd, "Public Displays of Connection," BT Technology Journal, vol. 22, no. 4, pp. 71–82, 2004.

[19] R. S. Burt, "The Social Capital of Structural Holes," in The New Economic Sociology: Developments in an Emerging Field, New York: Russell Sage Foundation., 2002, pp. 148–90.

[20] A. Portes, "Social Capital: Its Origins and Applications in Modern Sociology," Annu. Rev. Sociol., vol. 24, no. 1, pp. 1–24, Aug. 1998.

[21] P. S. Adler and S.-W. Kwon, "Social Capital: Prospects for a New Concept," The Academy of Management Review, vol. 27, no. 1, pp. 17–40, Jan. 2002.

[22] R. P. Putnam, "The Strange Disappearance of Civic America," The

American Prospect, vol. 24, 1996.

[23] R. P. Putnam, Bowling Alone. New York: Simon & Schuster, 2000.

[24] B. WELLMAN, A. Q. HAASE, J. WITTE, and K. HAMPTON, "Does the Internet Increase, Decrease, or Supplement Social Capital?," American Behavioral Scientist, vol. 45, no. 3, pp. 436 –455, Nov. 2001.

[25] B. Wellman and K. A. Frank, "Network capital in a multi-level world: Getting support from personal communities," in Social capital: Theory and research, 2001.

[26] Social behavior: its elementary forms: Homans, George Caspar, 1910-: Free Download & Streaming: Internet Archive. 2010.

[27] R. Kumar, J. Novak, and A. Tomkins, "Structure and evolution of online social networks," 2006, pp. 611–617.

[28] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee, "Measurement and analysis of online social networks," 2007, pp. 29–42.

[29] Y. Zhu, "Measurement and analysis of an online content voting network: a case study of Digg," Raleigh, North Carolina, USA, 2010, pp. 1039–1048.

[30] S. Barab, "Designing for virtual communities in the service of learning," The Information Society, Jan. 2003.

[31] B. Wellman and University of Toronto. Centre for Urban and Community Studies, Studying personal communities in East York / Barry Wellman. [Toronto]: Centre for Urban and Community Studies, University of Toronto, 1982.

[32] Q. Jones, "Virtual-Communities, Virtual Settlements & Cyber-Archaeology: A Theoretical Outline," Journal of Computer Mediated Communication, Jan. 1997.

[33] H. Rheingold, "A slice of my life in my virtual community," Global networks: Computers and international communication, pp. 57–80, Jan. 1993.

[34] B. Wellman, "Changing connectivity: A future history of Y2. 03K,"

socresonline.org.uk, Jan. 2000.

[35] J. Marathe, "Creating community online," Durlacher Research Ltd, 1999.

[36] L. Carotenuto, W. Etienne, M. Fontaine, J. Friedman, H. Newberg, M. Muller, M. Simpson, J. Slusher, and K. Stevenson, "Communityspace: Toward flexible support for voluntary knowledge communities," 1999.

[37] E. Wenger, "Communities of practice: Learning, meaning, and identity," 1999.

[38] T. Shummer, "Patterns for building communities in collaborative systems," Proceedings of the 9th European Conference on Pattern Languages and Programs, 2004.

[39] E. Wenger, R. McDermott, and W. Snyder, "Cultivating Communities of Practice: A Guide to Managing Knowledge," Harvard Business School Press, 2002.

[40] G. Probst and S. Borzillo, "Why communities of practice succeed and why they fail," European Management Journal, vol. 26, no. 5, pp. 335–347, 2008.

[41] A. Bourhis, L. Dubé, R. Jacob, and others, "The success of virtual communities of practice: The leadership factor," The Electronic Journal of Knowledge Management, vol. 3, no. 1, pp. 23–34, 2005.

[42] J. Plaskoff, "Intersubjectivity and community building: Learning to learn organizationally," in Collection, 2003, pp. 161–184.

[43] E. D. Mynatt, A. Adler, M. Ito, and V. L. O'Day, "Design for network communities," Atlanta, Georgia, United States, 1997, pp. 210–217.

[44] C. S. de Souza and J. Preece, "A framework for analyzing and understanding online communities," Interacting with Computers, vol. 16, no. 3, pp. 579–610, Jun. 2004.

[45] H. Saint-Onge and Debra Wallace Ph, Leveraging Communities of Practice for Strategic Advantage. Butterworth-Heinemann, 2002.

[46] L. Garton, C. Haythornthwaite, and B. Wellman, "Studying online

social networks," Journal of Computer-Mediated Communications, vol. 3, no. 1, 1997.

[47] P. Selznick, "In search of community," in Rooted in the land, 1996.

[48] E. Lesser and J. Storck, "Communities of practice and organizational performance," IBM Systems Journal, vol. 40, no. 4, pp. 831–841, 2001.

[49] A. Anderson, D. Huttenlocher, J. Kleinberg, and J. Leskovec, "Discovering Value from Community Activity on Focused Question Answering Sites: A Case Study of Stack Overflow," in Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, USA, 2012, pp. 850–858.

[50] S. Siersdorfer, S. Chelaru, J. S. Pedro, I. S. Altingovde, and W. Nejdl, "Analyzing and Mining Comments and Comment Ratings on the Social Web," ACM Trans. Web, vol. 8, no. 3, pp. 17:1–17:39, Jul. 2014.

[51] E. Kontopoulos, C. Berberidis, T. Dergiades, and N. Bassiliades, "Ontology-based sentiment analysis of twitter posts," Expert Systems with Applications, vol. 40, no. 10, pp. 4065–4074, Aug. 2013.

[52] C. M. Chiu, M. H. Hsu, and E. T. G. Wang, "Understanding knowledge sharing in virtual communities: an integration of social capital and social cognitive theories," Decision Support Systems, vol. 42, no. 3, pp. 1872–1888, 2006.

[53] K.-Y. Lin and H.-P. Lu, "Why people use social networking sites: An empirical study integrating network externalities and motivation theory," Computers in Human Behavior, vol. 27, no. 3, pp. 1152–1161, May 2011.

[54] Z. Yan, T. Wang, Y. Chen, and H. Zhang, "Knowledge sharing in online health communities: A social exchange theory perspective," Information & Management, vol. 53, no. 5, pp. 643–653, Jul. 2016.

[55] Z. Zhou, X.-L. Jin, and Y. Fang, "Moderating role of gender in the relationships between perceived benefits and satisfaction in social virtual world continuance," Decision Support Systems, vol. 65, pp. 69–79, Sep. 2014.

[56] D. Cartwright and F. Harary, "Structural balance: a generalization

of Heider's theory," Psychological Review, vol. 63, no. 5, pp. 277–293, 1956.

[57] C. Danescu-Niculescu-Mizil, G. Kossinets, J. Kleinberg, and L. Lee, "How Opinions Are Received by Online Communities: A Case Study on Amazon.Com Helpfulness Votes," in Proceedings of the 18th International Conference on World Wide Web, New York, NY, USA, 2009, pp. 141–150.

[58] N. Sun, P. P.-L. Rau, and L. Ma, "Understanding lurkers in online communities: A literature review," Computers in Human Behavior, vol. 38, pp. 110–117, Sep. 2014.

[59] H. Zhu, R. E. Kraut, and A. Kittur, "Effectiveness of Shared Leadership in Wikipedia," Human Factors: The Journal of the Human Factors and Ergonomics Society, vol. 55, no. 6, pp. 1021–1043, Dec. 2013.

[60] R. West, H. S. Paskov, J. Leskovec, and C. Potts, "Exploiting Social Network Structure for Person-to-Person Sentiment Analysis," arXiv:1409.2450 [physics], Sep. 2014.

[61] J. Cheng, C. Danescu-Niculescu-Mizil, and J. Leskovec, "How Community Feedback Shapes User Behavior," in Eighth International AAAI Conference on Weblogs and Social Media, 2014.

[62] P. Raeth, S. Smolnik, N. Urbach, and C. Zimmer, "Towards Assessing the Success of Social Software in Corporate Environments," AMCIS 2009 Proceedings, Jan. 2009.

[63] S. Fortunato, "Community detection in graphs," Physics Reports, vol. 486, no. 3–5, pp. 75–174, Feb. 2010.

[64] P. Kim and S. Kim, "Detecting overlapping and hierarchical communities in complex network using interaction-based edge clustering," Physica A: Statistical Mechanics and its Applications, vol. 417, pp. 46–56, Jan. 2015.

[65] J. Gairín-Sallán, D. Rodríguez-Gómez, and C. Armengol-Asparó, "Who exactly is the moderator? A consideration of online knowledge management network moderation in educational organisations," Computers & Education, vol. 55, no. 1, pp. 304–312, Aug. 2010.

[66] U. Matzat and G. Rooks, "Styles of moderation in online health and

support communities: An experimental comparison of their acceptance and effectiveness," Computers in Human Behavior, vol. 36, pp. 65–75, Jul. 2014.

[67] S. L. Toral, M. R. Martínez-Torres, and F. Barrero, "Analysis of virtual communities supporting OSS projects using social network analysis," Information and Software Technology, vol. 52, no. 3, pp. 296–303, Mar. 2010.

[68] S. A. Myers and J. Leskovec, "The Bursty Dynamics of the Twitter Information Network," in Proceedings of the 23rd International Conference on World Wide Web, New York, NY, USA, 2014, pp. 913–924.

[69] C. Lipizzi, L. Iandoli, and J. E. R. Marquez, "Combining structure, content and meaning in online social networks: The analysis of public's early reaction in social media to newly launched movies," Technological Forecasting and Social Change, vol. 109, pp. 35–49, Aug. 2016.

[70] A. Euerby and C. M. Burns, "Improving Social Connection Through a Communities-of-Practice-Inspired Cognitive Work Analysis Approach," Human Factors: The Journal of the Human Factors and Ergonomics Society, vol. 56, no. 2, pp. 361–383, Mar. 2014.

[71] L. Dubé, A. Bourhis, and R. Jacob, "Towards a typology of virtual communities of practice," Interdisciplinary Journal of Information, Knowledge, and Management, vol. 1, pp. 3–13, 2003.

[72] A. Hassan Zadeh and R. Sharda, "Modeling brand post popularity dynamics in online social networks," Decision Support Systems, vol. 65, pp. 59–68, Sep. 2014.

[73] X. Li, M. Wang, and T.-P. Liang, "A multi-theoretical kernel-based approach to social network-based recommendation," Decision Support Systems, vol. 65, pp. 95–104, Sep. 2014.

[74] R. Y. K. Lau, C. Li, and S. S. Y. Liao, "Social analytics: Learning fuzzy product ontologies for aspect-oriented sentiment analysis," Decision Support Systems, vol. 65, pp. 80–94, Sep. 2014.

[75] S. Amer-Yahia, S. Anjum, A. Ghenai, A. Siddique, S. Abbar, S. Madden, A. Marcus, and M. El-Haddad, "MAQSA: A System for Social

Analytics on News," in Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data, New York, NY, USA, 2012, pp. 653–656.

[76] R. Jones, S. Sharkey, J. Smithson, T. Ford, T. Emmens, E. Hewis, B. Sheaves, and C. Owens, "Using Metrics to Describe the Participative Stances of Members Within Discussion Forums," J Med Internet Res, vol. 13, no. 1, Jan. 2011.

[77] J. Lave and E. Wenger, "Situated Learning: Legitimate Peripheral Participation," books.google.com, Jan. 1991.

[78] E. M. Rogers and D. L. Kincaid, "Communication Networks: Toward a New Paradigm for Research," p. 386, 1981.

[79] B. Wellman and S. D. berkowitz, "Social Structures: A Network Approach," Emerald Group Publishing Limited, vol. 15, p. 528, 1998.

[80] R. L. Breiger, K. M. Carley, P. Pattison, and N. R. C. (U. S.). C. on H. Factors, "Dynamic Social Network Modeling and Analysis: workshop summary, Volume 2002," p. 379, Jan. 2003.

[81] S. Wasserman and K. Faust, "Social network analysis: Methods and applications," books.google.com, Jan. 1994.

[82] J. Xu and H. Chen, "CrimeNet explorer: a framework for criminal network knowledge discovery," ACM Transactions on Information Systems (TOIS), Jan. 2005.

[83] W. Huang, S.-H. Hong, and P. Eades, "How people read sociograms: a questionnaire study," Tokyo, Japan, 2006, pp. 199–206.

[84] J. Scott, "Social network analysis: a handbook" p. 208, Jan. 2000.

[85] J. Preece and D. Maloney-Krichmar, "Online Communities: Focusing on Sociability and Usability," Handbook of Human-Computer Interaction, Jan. 2003.

[86] D. Fetterman, "Ethnography: Step by Step," orton.catie.ac.cr, Jan. 1998.

[87] J. Nocera, "Ethnography and Hermeneutics in Cybercultural Research Accessing IRC Virtual Communities," Journal of Computer-Mediated

Communication, Jan. 2002.

[88] R. Kozinets, "The Field Behind the Screen: Using Netnography for Marketing Research in Online Communities," Journal of Marketing Research, Jan. 2002.

[89] L. Robinson and J. Schulz, "New field sites, new methods: new ethnographic opportunities."

[90] J. Koh, Y. Kim, B. Butler, and G. Bock, "Encouraging participation in virtual communities," Communications of the ACM, Jan. 2007.

[91] F. Sudweeks and S. J. Simoff, "Complementary explorative data analysis: the reconciliation of quantitative and qualitative principles," Doing Internet Research: Critical Issues and Methods for Examining the Net, pp. 29–55, 1999.

[92] A. Arenas, L. Danon, A. Diaz-Guilera, P. M. Gleiser, and R. Guimera, "Community analysis in social networks," The European Physical Journal B-Condensed Matter and Complex Systems, vol. 38, no. 2, pp. 373–380, 2004.

[93] K. Ehrlich, C. Lin, and V. Griffiths-Fisher, "Searching for experts in the enterprise: combining text and social network analysis," Proceedings of the 2007 international ACM conference on Supporting group work, Jan. 2007.

[94] M. Graves, A. Constabaris, and D. Brickley, "FOAF: Connecting People on the Semantic Web," Cataloging & Classification Quarterly, vol. 43, no. 3–4, pp. 191–202, Apr. 2007.

[95] J. G. Breslin, S. Decker, A. Harth, and U. Bojars, "SIOC: an approach to connect web-based communities," International Journal of Web Based Communities, vol. 2, no. 2, pp. 133–142, Jan. 2006.

[96] S. A. Ríos, F. Aguilera, F. Bustos, T. Omitola, and N. Shadbolt, "Leveraging social network analysis with topic models and the Semantic Web (extended)," Web Intelligence and Agent Systems: An International Journal, vol. 11, no. 4, pp. 303–314, 2013.

[97] S. A. Ríos, F. Aguilera, F. Bustos, T. Omitola, and N. Shadbolt,

"Leveraging social network analysis with topic models and the semantic web," in Web Intelligence and Intelligent Agent Technology (WI-IAT), 2011 IEEE/WIC/ACM International Conference on, 2011, vol. 3, pp. 339–342.

[98] A. J. Kim, Community Building on the Web: Secret Strategies for Successful Online Communities. Addison-Wesley Longman Publishing Co., Inc., 2000.

[99] J. Plaskoff, "Intersubjectivity and community building: Learning to learn organizationally," in The Blackwell handbook of organizational learning and knowledge management, Blackwell Publishing, 2003, pp. 161–184.

[100] F. Henri and B. Pudelko, "Understanding and analysing activity and learning in virtual communities," 2003. [Online]. Available: http://hal.archives-ouvertes.fr/hal-00190267/. [Accessed: 28-Apr-2010].

[101] L. Guo, E. Tan, S. Chen, X. Zhang, and Y. (Eric) Zhao, "Analyzing patterns of user content generation in online social networks," presented at the the 15th ACM SIGKDD international conference, Paris, France, 2009, p. 369.

[102] D. W. McMillan, "Sense of community," J. Community Psychol., vol. 24, no. 4, pp. 315–325, Oct. 1996.

[103] J. E. Puddifoot, "Dimensions of community identity," J. Community. Appl. Soc. Psychol., vol. 5, no. 5, pp. 357–370, Dec. 1995.

[104] D. Chavis, "Sense of Community Index."

[105] J. L. Nasar and D. A. Julian, "The Psychological Sense of Community in the Neighborhood," Journal of the American Planning Association, vol. 61, no. 2, pp. 178–184, Jun. 1995.

[106] A. L. Blanchard, "Testing a model of sense of virtual community," Computers in Human Behavior, vol. 24, no. 5, pp. 2107–2123, Sep. 2008.

[107] D. Chavis, K. Lee, and J. Acosta, "The sense of community (SCI) revised: The reliability and validity of the SCI-2," in 2nd international community psychology conference, Lisboa, Portugal, 2008.

[108] D. Blei, A. Ng, and M. Jordan, "Latent dirichlet allocation," The

Journal of Machine Learning, Jan. 2003.

[109] I. Bíró, D. Siklósi, J. Szabó, and A. A. Benczúr, "Linked latent Dirichlet allocation in web spam filtering," presented at the the 5th International Workshop, Madrid, Spain, 2009, p. 37.

[110] S. A. Ríos, J. D. Velásquez, H. Yasuda, and T. Aoki, "Web site offline structure reconfiguration: A web user browsing analysis," in International Conference on Knowledge-Based and Intelligent Information and Engineering Systems, 2006, pp. 371–378.

[111] S. A. Rıos, J. D. Velásquez, H. Yasuda, and T. Aoki, "Using a self organizing feature map for extracting representative web pages from a web site," International Journal of Computational Intelligence Research (IJCIR), vol. 2, pp. 159–167, 2006.

[112] S. A. Rios, J. D. Velasquez, E. S. Vera, H. Yasuda, and T. Aoki, "Establishing guidelines on how to improve the web site content based on the identification of representative pages," in The 2005 IEEE/WIC/ACM International Conference on Web Intelligence (WI'05), 2005, pp. 284–288.

[113] G. L'Huillier, S. A. Ríos, H. Alvarez, and F. Aguilera, "Topic-based social network analysis for virtual communities of interests in the Dark Web," presented at the ACM SIGKDD Workshop, Washington, D.C., 2010, pp. 1–9.

[114] I. Sato and H. Nakagawa, "Topic Models with Power-law Using Pitman-Yor Process," in Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, USA, 2010, pp. 673–682.

[115] R. A. Schwier, "Shaping the metaphor of community in online learning environments," 2002.