SISTEMA PARA IDENTIFICAR, EXTRAER Y VISUALIZAR CONOCIMIENTO EN CORREOS ELECTRÓNICOS

ING. JUAN DAVID CRUZ GÓMEZ

Tesis presentada como requisito parcial para obtener el título de

MSC. EN INGENIERÍA

INGENIERÍA DE SISTEMAS Y COMPUTACIÓN

Director:
ING. FABIO GONZÁLEZ OSORIO, PH.D.
Profesor Asociado

UNIVERSIDAD NACIONAL DE COLOMBIA
FACULTAD DE INGENIERÍA
BOGOTÁ D.C.
2008

Aprobada por la Facultad de Ingeniería en cumplimiento de los requisitos exigidos para otorgar el título de: **MSc. en Ingeniería — Ingeniería de Sistemas y Computación**

Ing. Fabio González Osorio, Ph.D.
Director de la Tesis

Ing. Denisse Cangrejo Aljure, Ph.D(c).
Jurado

Ing. Ingrid Paez, Ph.D.
Jurado

Jean Pierre Charalambos, Ph.D.
Jurado
Universidad Nacional de Colombia
Bogotá D.C., Noviembre de 2008

RESUMEN

SISTEMA PARA IDENTIFICAR, EXTRAER Y VISUALIZAR CONOCIMIENTO EN CORREOS ELECTRÓNICOS

por

ING. JUAN DAVID CRUZ GÓMEZ
MSc. en Ingeniería en Ingeniería de Sistemas y Computación

UNIVERSIDAD NACIONAL DE COLOMBIA

Director: Ing. Fabio González Osorio, Ph.D.

Este trabajo presenta un modelo de visualización de correos electrónicos y redes de contactos que le permiten al usuario realizar búsquedas y asociaciones basadas en la información no evidente de sus mensajes de correo electrónico. La información no evidente corresponde a toda aquella que se encuentra en las relaciones entre los contactos, en el cuerpo de los mensajes, los asuntos y los archivos adjuntos si estos existen.

El primer tipo de información muestra a personas y/u organizaciones involucradas de alguna forma con la persona dueña de los mensajes de correo. El segundo tipo de información muestra la distribución de temas que han sido tratados por las personas que aparecen en los mensajes de correo electrónico. Lo que se busca es relacionar los dos tipos de información y mostrarla de una forma que sea sencilla e intuitiva para el usuario.

Para esto se utiliza un modelo que dispone la información de temas en un mapa bi-dimensional y agrupa dichos temas en áreas, las cuales guardan en su disposición geométrica la relación entre ellas, es decir, entre más cerca estén, más relacionadas estarán. Luego, se relaciona la red de contactos identificada con el mapa de temas generado, y luego se utiliza un modelo basado en la fuerza de atracción que las diferentes áreas generan sobre cada uno de los nodos de la red de contactos, de modo que sea posible determinar la posición de cada uno de dichos nodos según la relación con los diferentes temas tratados.

De esta forma los dos tipos de información y sus representaciones son relacionados y mostrados en el mismo plano bi-dimensional, permitiendo identificar personas y los temas tratados por cada una de ellas de una forma intuitiva.

Durante el desarrollo del modelo, se planteó además un modelo de calificación de los algoritmos para pintar las redes de contactos en términos de la facilidad de interpretación para los usuarios y de la complejidad computacional inherente a una estructura compleja como lo son los grafos.

RECONOCIMIENTOS

Quiero hacer un reconocimiento especial a mi director de tesis, Fabio González y a todos los miembros del LISI por su apoyo, tanto en lo académico como en lo personal. Quiero agradecer también a mis compañeros de trabajo de la Universidad de los Andes y de la Pontificia Universidad Javeriana y ETB, en especial a Alberto García, Enrique González a José Roure y a Fabio Caicedo, que no solo me apoyaron durante el el tiempo que estuve allí, sino que me aportaron otros puntos de vista de diferentes temas y tópicos, los cuales me fueron útiles durante el desarrollo de este trabajo y seguramente lo serán durante mi posterior desarrollo profesional.

También quiero agradecer a Emanuela Todeva de Surrey University, Erwin Bendl de Siemens de Austria y a Bernie Hogan de Toronto University, aunque seguramente no entenderán este documento, por sus aportes y comentarios los cuales fueron también valiosos.

Finalmente, quiero agradecer a todas aquellas personas que dejé de mencionar aquí, no ha sido de forma voluntaria. Jamás habría podido hacer este pequeño aporte solo.

DEDICATORIA

A mis papás y a mi hermana, mi familia única y verdadera, de quienes siempre recibí apoyo y críticas constructivas para mi vida.

Contenido

Lista de Figuras

Lista de Tablas

Capítulo 1

Introducción

El correo electrónico se ha convertido en uno de los medios de comunicación más utilizados en todo el mundo debido a la facilidad que presenta para enviar mensajes e intercambiar ideas así como otro tipo de información con otras personas en cualquier parte del mundo en tan solo unos segundos. Esto hace que el volumen de información intercambiado por esta vía aumente a diario, especialmente cuando se tiene la posibilidad de agregar información en forma de archivos adjuntos.

Junto con el crecimiento del volumen de información aparece el problema de la gestión de los mensajes de correo electrónico: cada vez las personas tienen menos tiempo para leer, organizar y decidir que hacer con los mensajes recibidos.

Dado el volumen de información de los mensajes de correo electrónico de una persona, se puede pensar que el conjunto de mensajes contiene una cierta cantidad de conocimiento, el cual, no es fácilmente identificable, y que en cambio, puede resultar de utilidad para esa persona debido a que los mensajes pueden ser vistos como un registro histórico de información sobre diversos temas, que además, han sido tratados con diferentes personas, ya sea de forma directa o de forma indirecta.

Los mensajes de correo, tienen ciertas características que no son evidentes, como por ejemplo la red de contactos, la cual tiene el registro de las personas con las que se han establecido comunicaciones y está compuesta por el remitente y los destinatarios.

Adicionalmente a los contactos, un mensaje de correo contiene un mensaje, un asunto, y, en algunas ocasiones, un archivo adjunto. La combinación de estos elementos representa la idea o intención con la que el mensaje ha sido enviado. Esta información permite enriquecer la información contenida en un conjunto de mensajes de correo, al establecer las relaciones con otras personas y las características de estas relaciones, como los temas tratados, por ejemplo.

Como se mencionó anteriormente, esta información no es evidente, por lo que se hace necesario construir un sistema de análisis de correos electrónicos personales que permita identificar la estructura de las relaciones entre personas –redes sociales– así como los objetos, en torno a los cuales, dichas relaciones se concretan. Estos objetos pueden ser vistos como cualquier entidad en torno a la cual se desarrolla una idea, por ejemplo, un documento o un prototipo.

Para lograr dicha construcción es necesario realizar ciertos desarrollos que permitan trabajar directamente con los mensajes de correo electrónico de una forma organizada. El primer paso, consiste en proponer un esquema para la extracción y representación de información relevante (información de contactos, temas, palabras clave) en correos electrónicos y de los archivos adjuntos, luego es necesario desarrollar un módulo que permita construir redes sociales a partir de la información de contacto y temas de los correos electrónicos y desarrollar un esquema de identificación y asociación de los objetos intermediadores con las redes sociales construidas, luego se debe desarrollar un prototipo de software que permita extraer, visualizar y analizar la información de los correos y las redes contenidas en ellos, y finalmente, evaluar el desempeño del sistema desarrollado.

En este documento se plasma la forma en la que se construye el sistema antes mencionado. Para organizar la discusión, el documento esta organizado de la siguiente manera: en el Capítulo 2, se

presentan las bases teóricas y algunos trabajos anteriores en extracción de conocimiento y análisis de redes sociales. En el Capítulo 3, se plantea la forma en la que se construyen los diferentes componentes del sistema, así como la manera en la que ellos interactúan. En el Capítulo 4 se describen los métodos para identificar y construir las redes de contactos contenidas en el conjunto de mensajes de correo electrónico, así como diferentes esquemas de visualización de las mismas. En el Capítulo 5 se describen los métodos utilizados para identificar los temas y visualizar las áreas temáticas que los contienen. Adicionalmente en este capítulo se explica el procedimiento utilizado para unificar los esquemas de visualización de redes de contactos y temas. Finalmente, en el Capítulo 6, se presentan algunas conclusiones y propuesta de trabajo futuro.

1.1. Resumen de Objetivos y Resultados

Para este trabajo se propusieron una serie de objetivos: uno general y cinco específicos. Cada uno de los objetivos específicos estaba pensado para avanzar hacia el cumplimiento del objetivo general planteado, razón por la cual es importante resaltarlos en este espacio. Adicionalmente, se hará un resumen de los resultados obtenidos para cumplir dichos objetivos.

A continuación se presentan los objetivos y los resultados obtenidos:

1. **Proponer un esquema para la extracción y representación de información relevante (información de contactos, temas, palabras clave) en correos electrónicos:** se desarrolló un modelo que permitió identificar la red de contactos contenida en un conjunto de mensajes de correo electrónico, así como la información relevante requerida para categorizar y agrupar los mensajes del conjunto. Dicho modelo es descrito en la Sección 3.2.

2. **Desarrollar un módulo que permita construir redes sociales a partir de la información de contacto y temas de los correos electrónicos:** Se desarrolló una arquitectura que permite desarrollar experimentos de extracción de información, desde la definición del repositorio de mensajes hasta la obtención de la red social asociada al repositorio definido. Tanto el modelo como la descripción del prototipo y las pruebas realizadas se encuentran descritas en la Sección 4.1.

3. **Desarrollar un esquema de identificación y asociación de los objetos intermediadores con las redes sociales construidas:** Utilizando la parte de extracción de información del producto del objetivo **2**, se diseñó un modelo de categorización de información de los mensajes de correo electrónico con el fin de agrupar los mensajes y asignarles una localización geométrica en un área de dimensiones fijas. Tanto el modelo como la descripción del prototipo y las pruebas realizadas se encuentran descritas en la Sección 5.2.

4. **Desarrollar un prototipo de software que permita extraer, visualizar y analizar la información de los correos y las redes contenidas en ellas:** Se desarrolló un prototipo de software que permita implementar experimentos de identificación de redes de contactos y de grupos de mensajes utilizando los resultados de los objetivos **2** y **3** de modo que se muestre la relación entre la red de contactos y la información sobre los mensajes contenidos en un conjunto determinado. Tanto el modelo como la descripción del prototipo y las pruebas realizadas se encuentran descritas en la Sección 5.3.

5. **Evaluar el desempeño del sistema desarrollado:** Se desarrolló un modelos cualitativo para medir la eficiencia de los grafos generados en términos de la utilidad y la velocidad de pintado. Se supone que un esquema de visualización de información está diseñado para personas que interactúan con una máquina.

1.2. Publicaciones Asociadas al Desarrollo del Trabajo

Durante el desarrollo de este trabajo, y como producto asociado a él, se desarrollaron dos publicaciones, la primera, fué presentada en el Segundo Congreso Colombiano de Computación, y la segunda, la XXVII SUNBELT Conference.

A continuación de presentan las referencias de las publicaciones:

- C. Rodríguez, F. González, and J. D. Cruz, "Una aproximación a la categorización automática de correos electrónicos," in *Memorias del Segundo Congreso Colombiano de Computación*, (Bogotá, Colombia), Pontificia Univerisdad Javeriana, Abril 2007.
- J. D. Cruz and F. González, "A social network based model for e-mail information visualization," in *Proceedings of the XXVII SUNBELT Conference*, (Corfu, Greece), International Network for Social Networks Analysis, INSNA, May 2007.

Es importante resaltar que la conferencia SUNBELT es patrocinada por el *International Network for Social Network Analysis* – INSNA, y es una de las conferncias más representativas en el área del análisis de redes sociales.

Capítulo 2

Redes Sociales y Extracción de Conocimiento de Correos Electrónicos

> "What Descartes did was a good step. You have added much several ways, & especially in taking ye colours of thin plates into philosophical consideration. If I have seen further it is by standing on ye shoulders of Giants." –Newton to Hooke, 5 Feb. 1676.

Como se ha mencionado antes, los correos electrónicos están compuestos de dos tipos de información, cuyos vínculos no son evidentes para los usuarios, sin embargo, pueden resultar de gran utilidad para ellos.

Actualmente existen diversos estudios referentes al manejo de las redes sociales y la manipulación de textos con el fin de extraer información de ellos, y, aunque de han hecho avances interesantes, no es mucho lo que se ha hecho por intentar unificar los dos tipos de información y utilizar un modelo de visualización sencillo pero efectivo para presentar a un usuario.

En este capítulo, se presentan algunas bases teóricas sobre el análisis de redes sociales y extracción de conocimiento de textos y específicamente, de correos electrónicos.

2.1. Redes Sociales

Una red es un conjunto de elementos de dos tipos, vértices o nodos y arcos que representan una conexión entre ellos. En general, pueden ser vistos como un grafo, sobre el cual se pueden aplicar diversas operaciones. En la teoría de redes sociales, los nodos, representan *actores*, los cuales pueden ser individuos u organizaciones, en general, entidades sociales y los arcos, representan las *relaciones* entre dichos actores.

En la teoría de redes sociales existen varios conceptos ampliamente utilizados. Wasserman [1] explica dichos conceptos, los cuales se resumen a continuación:

- **Actor:** Un actor es una entidad social, que pueden ser personas, organizaciones, asociaciones o naciones. Aunque se utiliza el término actor, no necesariamente actúa, sino que tiene la habilidad de comunicarse con otros actores de su misma clase.
- **Vínculo Relacional:** Estos son los vínculos que encadenan a los actores. En el análisis de redes sociales existen diversos tipos de vínculos, algunos de ellos son: los que reflejan amistad, transferencia de recursos, tangibles o no, asociación o afiliaciones, conexiones físicas, entre otras.
- **Diada:** Es la representación de los posibles vínculos entre dos actores de una red. Su importancia radica en el hecho de ser la unidad básica que se puede estudiar en una red social.
- **Triada:** Es la representación de una subred de tres actores y las posibles relaciones que pueden existir entre ellos.

- **Subgrupo:** Un subgrupo consiste en un subconjunto de actores y las relaciones entre ellos. En general, un subgrupo debe ser estudiado con algún criterio de medida.
- **Grupo:** Un grupo está conformado por todos los actores cuyos vínculos van a ser medidos.
- **Relación:** Es la colección de vínculos del mismo tipo entre los miembros de un grupo. Por ejemplo los vínculos de amistad, o las relaciones diplomáticas entre naciones.

El análisis de redes sociales va más allá de simplemente estudiar desde la perspectiva matemática un grafo, sino que tiene en cuenta los elementos personales y sociales involucrados en las redes. Es por esto que se centra principalmente en el estudio del flujo de información y recursos entre los actores de la red [1].

Las redes sociales son representaciones de las interacciones entre personas o entre organizaciones. La representación es forma de un grafo dirigido, en donde los nodos corresponden a individuos (personas, organizaciones o cualquier asociación) y los arcos son los vínculos entre ellos. La teoría que rodea a las redes sociales es utilizada para analizar los vínculos entre individuos, descubrir posibles fallas de comunicación (cuellos de botella) o redes que presentan un peligro potencial al tener individuos que puedan desconectarla (generar huecos sociales) y generar dos o mas sub-redes, también es utilizada para gestionar el capital intelectual de la organización, por ejemplo, descubrir quién sabe que.

2.1.1. Tipos de Redes Sociales

Las redes sociales tiene características en su estructura que varían según la finalidad que tengan. Características como los tipos de nodos que pueden tener y la estructura, pueden cambiar desde la forma en que se representan, hasta la forma en que el análisis es llevado a cabo.

Existen tres tipos básicos de red social: unimodal, bimodal y egocéntrica [1] las cuales de presentan a continuación.

2.1.1.1. Red unimodal

Una red unimodal es aquella cuyos actores pertenecen al mismo tipo, es decir, solo existe un conjunto de actores. Las relaciones entre actores en este tipo de red, representan vínculos específicos entre dichos actores. Por ejemplo, se tienen vínculos de amistad, transferencia de objetos, tangibles o intangibles, compras, ventas, interacción, movimientos, cambios de lugar, roles y hasta vínculos de parentesco.

2.1.1.2. Red bimodal

En una red bimodal existen dos tipos de actores, los cuales están relacionados entre si a través de vínculos de afiliación. Los dos tipos de actores pueden ser organizaciones y personas, cuyos vínculos están representados por la pertenencia a determinada organización

2.1.1.3. Red egocéntrica

La principal característica de una red egocéntrica es que existe un actor que tiene un grado de intermediación muy alto. Esto quiere decir que uno de los nodos concentra el mayor número de conexiones que otros en la misma red. En general, este tipo de redes son las que representan las conexiones entre los contactos de correo electrónico de una persona.

2.1.2. Generación

Existen dos formas de generar las redes sociales, la primera, es a través de encuestas y entrevistas. Esta forma es muy utilizada por las empresas de consultoría ya que permite extraer de forma más directa el conocimiento tácito de las relaciones existentes. Este tipo de encuestas se realiza en torno

a algún tema, es decir, hay preguntas del tipo "cuando pasa el evento x, ¿a quién le informa?" lo que hace que se generen nodos o personas que pueden estar dentro o fuera de la organización y al final, cuando las entrevistas terminan, se llena una matriz de adyacencias con la que se genera el grafo. Esta forma, aunque es muy dispendiosa, es bastante conveniente para empresas que están en un proceso de implementación de un programa de gestión de conocimiento y deben analizar los caminos de comunicación existentes.

Sin embargo, dentro de las empresas, existe otra forma de generar las redes. Esta forma es automática y utiliza la información encontrada en los correos corporativos, lo que resulta natural, ya que los correos contienen un individuo de origen o remitente, un individuo (o muchos) destinatarios y un tema que los relaciona. El principal inconveniente de esta técnica es que necesita obtener información de los correos electrónicos de cada persona en la compañía, lo que puede generar incomodidades en los empleados ya que pueden sentir esto como una intrusión.

Adicionalmente, otra forma de generar redes sociales consiste en extraer la información de citación de artículos científicos. Esta técnica no es muy útil para una empresa, o al menos para una en la que no hayan publicaciones, pero si es muy útil para encontrar redes de colaboración para crear documentos y descubrir expertos en diferentes áreas.

2.1.2.1. Visualización de Redes Sociales

En general la visualización se refiere a cualquier técnica para crear imágenes, diagramas y representaciones para comunicar un mensaje determinado. Las técnicas de visualización han sido desarrolladas a través de la historia de la humanidad para comunicar tanto ideas concretas como abstractas. Con el apoyo de las ciencias de la computación, las técnicas de visualización han evolucionado y permiten plasmar diferentes tipos de información y conocimiento orientadas al apoyo en la toma de decisiones y el diseño de diferentes elementos tecnológicos.

Debido a que la visualización está orientada a las personas, es necesario dotarla de elementos objetivos que permitan que el resultado mostrado sea entendido por un grupo particular de personas. Esto quiere decir que las técnicas deben permitir que a partir de un conjunto de datos e información se pueda crear un modelo entendible e intuitivo para un grupo de personas que se supone tienen ciertas habilidades para interpretar la información mostrada. La tarea de crear una técnica intuitiva no es sencilla debido a las interpretaciones que se le pueden dar a los datos y la forma en que ellos pueden ser mostrados.

En este caso se pretende desarrollar un modelo de representación de la información contenida en un conjunto de correos electrónicos utilizando las redes sociales contenidas en dicho conjunto así como en los temas que lo componen.

El primer componente se refiere a la visualización de las redes de contactos, el cual se construye la matriz de adyacencias a partir de los mensajes de correo y con base en ella dibuja el grafo. Para ello, existen diversos esquemas de organización de los nodos. La mayoría de ellos utilizan algoritmos de optimización, basados en las distancias y en los grados de la red social, sin embargo, tienen complejidades altas.

Uno de los esquemas más utilizados es el de Fruchterman & Reingold, presentado en 1991 por Fruchterman [2]. Este realiza una organización interesante del grafo, pero tiene una complejidad $O\left(n^2\right)$, por lo que no es recomendable para grafos con demasiados nodos.

En la Figura 2.1 se muestra la representación de la red de contactos de un conjunto de mensajes de correo organizados utilizando el algoritmo de Fruchterman & Reingold.

Adicionalmente, se implementó el algoritmo de escalamiento multidimensional (MDS por sus iniciales en inglés), el cual, es una técnica que permite, a partir de una matriz de distancias, generar puntos en espacio $n-$dimensionales que las representan. En este caso, las dimensiones son dos y se combinan dos técnicas, MDS métrico, o clásico, y MDS no métrico, que se basa en un método de optimización de ascenso en gradiente.

Mayores detalles sobre el MDS tanto métrico como no métrico [3].

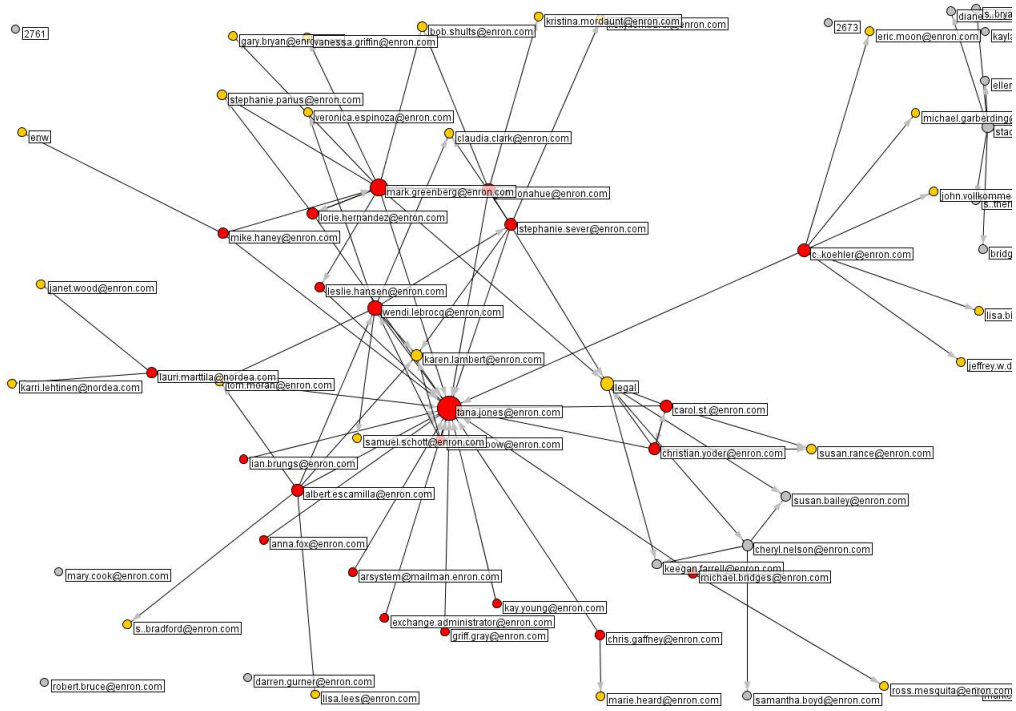


Figura 2.1: Resultado del algoritmo de Fruchterman & Reingold

En la Figura 2.2, se puede ver como queda organizado el grafo cuando se utiliza escalamiento multidimensional.

2.1.3. Análisis de Redes Sociales

Una vez se ha construido la red social, es necesario analizarla ya que la red por si sola no aporta mucha información. La teoría de redes sociales está fundamentada en la teoría de grafos, por lo que en general se pueden utilizar todas la medidas asociadas a esta última. En el análisis de redes sociales, se pueden clasificar en tres tipos de medidas: medidas de relación, medidas para individuos y medidas generales. A continuación se explican dichas medidas.

- Medidas Individuales: estas medidas permiten estudiar a cada uno de los miembros de la red de forma particular, por ejemplo, establecer que tanto participan en el flujo de información establecido sobre algún tema.

  - Grado: Número de vínculos directos con otros individuos
    - Grado de entrada: Número de vínculos al individuo desde otros individuos.
    - Grado de salida: Número de vínculos del individuo hacia otros individuos.
  - Intermediación: grado en el que un individuo es intermediario entre otros dos en el camino más corto entre ellos.
  - Cercanía: Grado en el que un individuo está, o puede alcanzar más fácilmente a otros individuos en la red.
  - Centralidad: Grado en el que un individuo es central en la red.

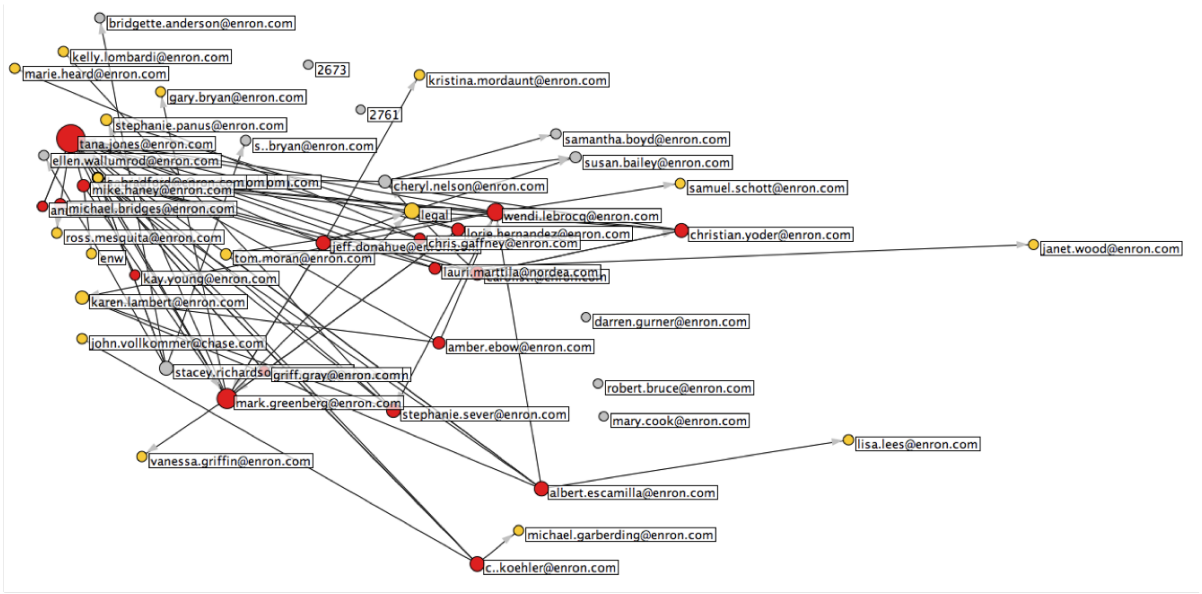


Figura 2.2: Experimento utilizando multidimensional scaling.

  - Prestigio: Los individuos prestigiosos son aquellos que son más fuentes de vínculos.
  - Rango: Número de vínculos con otros, donde otros son individuos que son de otro grupo o tienen otro estatus.

- Medidas de vínculos: estas medidas proveen información acerca del flujo de información en la red, como se genera y como se acaba.

  - Vínculos indirectos: Caminos con otros individuos en donde tales rutas contienen uno o más individuos en el medio.
  - Frecuencia: Cuantas veces ocurre el vínculo.
  - Estabilidad: Existencia del vinculo en el tiempo.
  - Multiplicidad: Grado en el que dos individuos están vinculados por diferentes caminos.
  - Fuerza: Cantidad de tiempo, intensidad emocional y reciprocidad en la presentación de servicios de dos individuos.
  - Dirección: Grado en el que un vínculo determinado va desde un individuo a otro.
  - Simetría: Grado en el que un vínculo es bi-direccional.

- Medidas generales: estas medidas permiten ver y analizar la red como un todo, por ejemplo, medir la penetración en una empresa de ciertos temas que pueden resultar estratégicos.

  - Tamaño: Número de individuos en la red.
  - Inclusividad: Número total de actores en la red menos el número de actores desconectados o aislados.
  - Conectividad: Grado en el que los individuos de la red están conectados con otros, ya sea directa o indirectamente.
  - Densidad: Proporción de los vínculos actuales con respecto a todos los posibles vínculos existentes.

- Simetría: Proporción de vínculos simétricos con respecto al número total de vínculos en la red.
- Transitividad: Número de caminos de tamaño 2.

En la práctica, no son utilizadas todas las medidas mencionadas anteriormente, en general, las más utilizadas son:

- Grado: Un individuo con alto grado, tiene más posibilidades de satisfacer sus necesidades y se reduce la dependencia con otros individuos.
- Intermediación: Un individuo con alto grado de intermediación, se convierte en indispensable y al mismo tiempo se convierte en un punto vulnerable de la red, ya que puede generar huecos estructurales que desconectan otros individuos o redes completas.
- Cercanía: Los individuos con un alto grado de cercanía, tienen a su disposición los caminos más cortos hacia otros individuos, por lo que se encuentran en una excelente posición para controlar el flujo de información que fluye a través de esta.

Adicionalmente, las redes sociales son objeto de otro tipo de análisis en los que se buscan huecos estructurales los cuales pueden generar desconexiones y aislamientos de porciones de la red. En [4], se trata este aspecto y se realiza un análisis y se desarrolla una herramienta capaz de encontrar este tipo de fallas en una red.

2.1.4. Aplicaciones de las Redes Sociales

Las redes sociales y su análisis tienen varias aplicaciones. Las primeras se refieren a la gestión del capital intelectual en forma de conocimiento tácito a través del mejoramiento de la comunicación entre los empleados de una compañía de modo que se puedan establecer redes de colaboración y de esta forma mejorar el desempeño corporativo así como para descubrir expertos en temas específicos.

Recientemente, se han utilizado estas técnicas para analizar la información de correos (a nivel personal) de modo que se puedan clasificar y realizar diferentes tareas con ellos, todo encaminado a automatizar tareas diarias que consumen bastante tiempo.

2.1.4.1. Trabajo colaborativo

Hoy en día, diferentes empresas trabajan en diferentes proyectos que requieren en muchos casos expertos en ciertas áreas del conocimiento que resuelvan dudas o problemas complejos. Cuando los problemas o las dudas no son muy grandes, se puede utilizar un buscador en Internet para encontrar la respuesta al problema, pero en muchos casos, se gasta mucho tiempo buscando, y en algunos casos menos afortunados, no se encuentra una respuesta satisfactoria.

Cuando una compañía es lo suficientemente grande como para que todos sus empleados no se conozcan entre si, es posible que una persona tenga la respuesta a una pregunta o sepa resolver un problema de otro empleado sin que este último sepa de la existencia del otro. Aquí es en donde entran las redes sociales.

A través de un análisis de estas se pueden identificar expertos en diferentes áreas dentro, y en algunos casos, fuera de la compañía. Esta identificación se puede hacer de forma manual, en la que el empleado debe averiguar el nombre de la persona que busca y escribirle un correo o llegar a ella a través del camino señalado en la red. Otra forma de identificación es a través de un sistema automático de recomendación, como se plantea en [5].

En [6] una organización es vista como una red actores en la que existen relaciones del tipo *quién quiere qué con quienes.* Para responder esta clase de preguntas, utilizan el análisis de redes sociales definiendo una serie de parámetros según los cuales se pueden agrupar los actores dentro de la red y miden el impacto de estos dentro de la organización.

Otra aplicación que resulta útil, desde el punto de vista empresarial y de gestión del conocimiento corporativo, es la identificación de líderes dentro de grupos. Estos grupos son en general formados durante la ejecución de proyectos que pueden o no, ser estratégicos para la compañía. En el trabajo desarrollado por Carvalho [7] se utilizan técnicas estadísticas a través de las cuales se estudian las formas en que cada una de las personas en la red envían los mensajes. Las formas en que pueden ser enviados los mensajes han sido agrupados de dos maneras: *broadcast* y no *broadcast*. El primero se refiere a los mensajes enviados por una persona del grupo al resto de personas del grupo, el segundo, se refiere a los mensajes que son enviados por una persona del grupo a otra persona del grupo. El estudio es llevado a cabo estudiando el flujo de mensajes dentro de la red social que se construye entorno a un tema, el cual es generalmente un proyecto corporativo.

Un trabajo similar al anterior es el desarrollado por Johansen [8]. En este caso la idea es encontrar, a través del análisis del tráfico de mensajes en la red social, comunidades de práctica en las compañías.

2.1.4.2. Gestión de correos

El correo electrónico es hoy en día una de las herramientas de comunicación mas utilizadas, y así mismo, el volumen de información recibido por este medio es muy grande, en algunos casos es tan grande, que las personas no tienen tiempo para responder los mensajes que le llegan. Sumado a lo anterior está el *spam* o correo no solicitado, el cual añade más trabajo a la revisión diaria de correos.

Para resolver estos problemas se han desarrollado diferentes técnicas, que permiten resolver de forma automática dichos problemas.

Neustaedter [9] propone una herramienta para decidir sobre que hacer con el correo entrante basado en la red social cercana al destinatario. Si el remitente no pertenece a su red, el correo es catalogado como spam, de otra forma es depositado en una carpeta apropiada. Otra herramienta es propuesta por Golbeck [10], en la que se decide si un correo es clasificado como no solicitado según la reputación que tenga en remitente en la red del destinatario.

En el trabajo de Khoussainov [11] se propone una técnica de gestión de correos en la que se relacionan los correos entrantes y los temas de los que trata cada uno. La idea propuesta es la de realizar un análisis de los mensaje entrantes y relacionarlos para clasificarlos. Aunque no hay individuos involucrados, si se genera una red de relaciones entre documentos.

Smith [12] propone el desarrollo de un sistema que permita decidir en que momento y a quién se le puede reenviar un correo electrónico basado en la red social actual. Esto permite que el trabajo sea delegado de forma más rápida y además permite que el conocimiento sobre algún tema se difunda en la red actual, y en algunos casos, que la red actual crezca añadiendo actores. Para esto se propone utilizar la herramienta propuesta por Neustaedter [9].

En el trabajo de Surendran [13], proponen realizar un análisis del comportamiento de los mensajes de salida de un usuario determinado con el fin de realizar una clasificación de sus mensajes. Esta clasificación está orientada a evitar mensajes de publicidad no deseada y virus que se difunden como gusanos en los mensajes.

Todas las técnicas anteriores utilizan aproximaciones estadísticas o que utilizan información contenida en las redes sociales generadas por los mensajes para gestionar el correo. Sin embargo, existen otras aproximaciones, como la propuesta por Secker [14], la cual utiliza sistemas inmunológicos artificiales. La idea fundamental es dividir los mensajes de correo en dos categorías: los interesantes y los que no lo son basado en la experiencia previa extraída del comportamiento del usuario. Aunque esta última aproximación no tiene en cuenta de forma explícita la red social del usuario, es una forma interesante y novedosa de gestión de mensajes.

En 2006, Minkov [15], propone un *framework* para crear grafos a partir de la información contenida en correos electrónicos. Estos grafos no son vistos como una red social común, sino que cada nodo y cada arco tiene propiedades diferentes, creando de esta forma un grafo que es bastante complejo de analizar desde la perspectiva del análisis de redes sociales. Su principal aplicación es la generación automática de reuniones, es decir, la elección de personas para una reunión determinada.

2.1.4.3. Otras aplicaciones

Una de las aplicaciones para las cuales fueron pensadas las redes sociales es la de gestión de conocimiento tácito a través de la medición del flujo de información entre individuos. Sin embargo, diversas aplicaciones derivadas de la gestión de conocimiento y que basan su funcionamiento en las redes sociales y su análisis respectivo han surgido recientemente. La mayoría de estas aplicaciones están orientadas a los negocios y su relación con sus clientes y buscan entregar a estos últimos el mayor valor con el fin de mejorar el posicionamiento de los primeros.

Una de estas aplicaciones es obtener información de la red de valor de los clientes, es propuesta por Domingos [16], y lo que hace es representar los mercados como redes sociales. Dentro de los mercados están incluidos los clientes y los productos; se asume que cada cliente puede o no, comprar un producto determinado y además puede influenciar la compra de este u otros productos a sus vecinos cercanos. La idea es estudiar la red generada por estas relaciones de influencia y obtener provecho de mercadeo.

En 2002, Chakrabarti [17], propone una aplicación que permita estudiar la estructura de tópicos amplios en la Web. En este caso se aplican algunas de las medidas del análisis de redes sociales para realizar el estudio, lo cual permite ver el flujo de la información entre comunidades, estados y países, encontrando así la denominada geografía de la Web. En el trabajo de Amento [18], se plantea un sistema que, basado en la información de páginas y sitios Web permite organizar tópicos sobre temas específicos lo que permite a los usuarios gastar menos tiempo en sus búsquedas. La solución propuesta utiliza técnicas basadas en el análisis de redes sociales y sistemas de recomendación para esto.

En [19] se propone un estudio acerca de la centralidad y el desempeño de individuos en comunidades de investigación y desarrollo. La hipótesis en la que se basa este estudio es que el rol funcional, el estatus y la forma en que se comunica con otros, tienen una influencia directa sobre un individuo y una influencia indirecta sobre su centralidad. Para este análisis utilizan técnicas derivadas del análisis de redes sociales y un conjunto de mensajes de correo electrónico dividido en dos periodos separados cuatro años.

Liben–Nowell [20] plantea una pregunta: "dada una red social en un tiempo t, como se puede predecir de forma exacta, los vértices que aparecerán antes de llegar al tiempo t'". Esto es, predecir la evolución de una red social en un contexto determinado, como redes de coautoría, teniendo en cuenta factores externos y propios de la red. Mirza [21] propone un sistema de recomendación basado en el análisis de grafos y en los saltos que existen entre individuos. Este trabajo resulta interesante desde el análisis de redes sociales dado que las redes sociales son representadas como grafos dirigidos.

O'Donell [22], plantea la utilización de las redes sociales extraídas de los mensajes de correo electrónico, detectar comportamientos no autorizados. Para esto realizan análisis estadísticos de los grafos luego de definir políticas sobre la red.

Otra aplicación, ampliamente utilizada, es la del apoyo a la gestión y transferencia de conocimiento en las organizaciones. Una buena descripción del proceso de transferencia de conocimiento y el apoyo que le dan las redes sociales es presentada en el trabajo de [23].

2.2. Procesamiento de Correo Electrónico

Los correos electrónicos son utilizados para enviar diversos tipos de información, que van desde invitaciones hasta informaciones importantes en el día a día del desarrollo de un proyecto. Aunque en general los correos pueden ser tratados como documentos genéricos, en realidad no poseen la estructura de estos últimos; esto se debe a que, en general, los correos electrónicos son utilizados para enviar mensajes no muy extensos y que además están inmersos en un contexto que las personas involucradas conocen. El contexto se refiere al marco en el que un mensaje entre dos o mas personas es generado y que hace que tal mensaje no requiere de una introducción y en algunos, que se espere una respuesta no explícita.

Dentro de tales contextos se pueden incluir reuniones, proyectos, invitaciones y otros temas de índole personal. Por ejemplo, durante la ejecución de un proyecto es necesario establecer un conjunto de reuniones las cuales hacen parte del contexto "Proyecto". En este contexto particular, un mensaje

de invitación a una reunión o de recordatorio de asistencia, al estar dentro dicho contexto, no necesitan de una introducción o una explicación, sino que con un mensaje de una línea basta. Otro ejemplo es presentado por Kusui [24] en donde se muestra una aplicación de gestión de conocimiento a partir del estudio de la estructura del correo y del hilo que se forma en una conversación.

Sin embargo, en muchas aplicaciones de extracción de información de correos, se asume cierta estructura que concuerda con la información buscada. Por esta razón, se explicarán algunas técnicas para el tratamiento de documentos y adicionalmente algunas aplicaciones de extracción de conocimiento de correos electrónicos con estas técnicas.

Las técnicas de extracción de conocimiento han sido desarrolladas de diferentes formas, algunas, con técnicas determinísticas, otras probabilísticas y otras basadas en técnicas bio-inspiradas, pero de forma más general, las técnicas se clasifican en aquellas que tienen aprendizaje supervisado y aquellas que tienen aprendizaje no supervisado, las primeras son utilizadas también para clasificar y las segundas para realizar tareas de agrupamiento. A continuación se explican algunas de ellas.

2.2.1. Representación de Correos para Extracción de Conocimiento

El correo electrónico, en general, puede ser tratado como texto, el cual tiene algunas características particulares como las explicadas al inicio de la sección, sin embargo, es posible utilizar algunas de las técnicas disponibles para el procesamiento de textos. Dichas técnicas buscan "vectorizar" el texto a ser estudiado. La técnica utilizada es encontrar la frecuencia de cada una de las palabras en un texto y cada una de dichas frecuencias corresponderá a un valor, normalizado, en el eje determinado por dicha palabra. Es decir, que si un documento contiene n palabras diferentes, el espacio generado es $n-$dimensional, haciendo que con algunas técnicas el tratamiento de los documentos sea complejo. Otro posible problema de esta representación es que todas las palabras tienen la misma importancia o peso, por ejemplo, en un texto sobre cálculo, la palabra derivada tiene la misma importancia que la, lo cual no es necesariamente cierto. Es por esto que se sugieren algunas modificaciones, como añadir una lista de palabras excluidas o *stop list*, en la que se añaden cosas como artículos y otros conectores. Esto es conocido como bolsa de palabras.

Otra modificación consiste en dar a cada palabra un peso dentro del documento. Para esto en [25] se propone un método general, el cual es base para otros desarrollos y el cual es llamado frecuencia inversa del documento (IDF por sus iniciales en inglés).

Black [26] utiliza la técnica IDF para extraer información sobre invitaciones en correo electrónico. En este caso se busca una estructura específica en un mensaje de correo, el cual debe tener la invitación a un evento, ya sea de manera formal o informal. El resto de correos no son tenidos en cuenta.

2.2.2. Extracción de Conocimiento

Los documentos y textos en general, poseen una gran cantidad de información que no es evidente para sus dueños. Dicha información no evidente, es por ejemplo, la relación entre un conjunto de documentos separados.

Para identificar esa información no evidente se han desarrollado un conjunto de técnicas de identificación y extracción de conocimiento. Dichas técnicas pueden ser basadas en técnicas supervisadas, no supervisadas y otras técnicas.

A continuación se explican con mas detalle cada una de ellas.

2.2.2.1. Análisis Supervisado

Las técnicas de aprendizaje supervisado buscan que los modelos sean obtenidos a partir de la comparación de la salida del procesamiento, hecho a partir de un conjunto de patrones de entrada, con un conjunto de patrones de salida. Tal comparación permite encontrar un error que permite recalcular los parámetros con los cuales fue hecho el procesamiento y buscar reducir tal error. Algunas de las técnicas basadas en tal análisis se resumen a continuación.

- **Modelos ocultos de Markov:** Los modelos ocultos de Markov (HMM por sus iniciales en inglés) son una herramienta utilizada para modelar procesos que varían en el tiempo. Están compuesto de nodos o estados, los cuales tienen una distribución de probabilidad de emisión de símbolos y una probabilidad de transición entre estados. En general, un modelo oculto de Markov queda totalmente especificado utilizando los siguiente elementos [27]: N, el número de estados del modelo, M, el número de símbolos que puede emitir un estado en un momento determinado, $\mathbf{P} = \{p_{ij}\}$, la matriz de probabilidad de transición de estados (probabilidad de ir del estado i hacia el estado j), $\mathbf{B} = \{b_j\,(k)\}$, un vector con las probabilidades de emitir el símbolo k en el estado j y la distribución de probabilidad del estado inicial.
  Aprovechando el hecho de que en un modelo de Markov solo interesa el valor del estado inmediatamente anterior, el modelo oculto de Markov se entrena para encontrar la matriz $\mathbf{P}$ y el vector$\mathbf{B}$, de modo que se pueda encontrar el camino más probable entre los estados dado un símbolo de entrada. Detalles sobre los diferentes métodos de entrenamiento pueden ser encontrados en [27]. En el caso de la clasificación de documentos, cada uno de los estados o clases está en capacidad de emitir uno de los términos del vocabulario con cierta probabilidad. Una aplicación de este concepto es mostrada por Denoyer [28]. En el caso de extracción de información de correos, Sarawagi [29], utiliza una modificación de los modelos ocultos de Markov para encontrar información relevante. Esta información también debe coincidir con una estructura determinada.

- **Redes neuronales:** Las redes neuronales realizan una simulación de lo que se cree sucede en el cerebro, en donde una neurona emite o no una señal dependiendo del nivel de las señales de entrada que recibe, es decir, si las señales de entrada superan cierto umbral. Con esta idea se construyen las neuronas artificiales, las que luego son conectadas para formar las llamadas redes neuronales.
  En las redes neuronales artificiales se busca adaptar los pesos de las conexiones entre las diferentes capas de neuronas para encontrar una función que modele el conjunto de datos de entrada. El entrenamiento de las redes neuronales artificiales puede ser supervisado o no supervisado, lo cual está directamente ligado al algoritmo de entrenamiento y a la topología de la red. En el primer tipo de entrenamiento se requiere que para cada patrón de entrada exista uno de salida, esto con el fin de compara la salida obtenida y calcular el error. Las generalidades sobre la arquitectura, tipos de unidades y entrenamiento de redes neuronales no están dentro del alcance de este trabajo, sin embargo información interesante puede ser encontrada en: [30] y [31]. Adicionalmente, se pueden encontrar algunos algoritmos de entrenamiento supervisado y no supervisado en [32] .

- **Sistemas inmunológicos artificiales:** Los sistemas inmunológicos artificiales son sistemas que intentan simular el comportamiento de los sistemas de defensa corporales. En el cuerpo, la idea es detectar todo aquello que no pertenece al cuerpo y eliminarlo. Existen varias teorías biológicas sobre la forma en que el cuerpo realiza esta tarea, pero no hay nada comprobado realmente, sin embargo, diferentes personas han comenzado a modelar los procesos descritos por dichas teorías, entre otros, la selección negativa propuesta por Forrest [33] o la selección clonal discutida en [34].
  En general, las aplicaciones de estas técnicas están dirigidas hacia la clasificación y el agrupamiento de diferentes tipos de datos, pero debido a la inspiración de la cual provienen, las clasificaciones son propio o extraño. Este hecho, de utilizar únicamente dos clases para clasificar podrá suponer un inconveniente a la hora de clasificar documentos ya que pueden existir bastantes clases de ellos. Lo que se hace entonces es reclasificar el conjunto de entrenamiento para que solo tenga dos clases, por ejemplo, Twycross [35] propone que los documentos sean relevantes o irrelevantes. Con esta división ya es posible comenzar a entrenar el conjunto de detectores, los cuales son los encargados de realizar la clasificación, los cuales, pueden ser generados utilizando selección negativa, selección positiva o alguna aproximación evolutiva como en [35]. Una vez entrenados los detectores, el sistema está en capacidad de clasificar cualquier documento nuevo que se le presente. En los casos anteriores se utilizan generalmente dos clases, de modo que se

sigue el modelo propio - extraño, sin embargo, Marwah [36] propone una pequeña modificación que se basa en la afinidad de los detectores con los antígenos, en donde, entre más afín sea un antígeno con un detector determinado, querrá decir que más se parece a la clase que representa.

- **Computación evolutiva:** La computación evolutiva recoge varios métodos de optimización que basan su funcionamiento en la teoría de la evolución natural de Darwin y en los conceptos y términos de la genética. De esta forma, dentro de las técnicas evolutivas se habla de poblaciones, individuos, cromosomas, genes y funciones de adaptación, lo cual es fundamental a la hora de establecer la representación para un problema particular.
  Un trabajo interesante es el propuesto por Picarougne [37], en el que plantea la utilización de un algoritmo genético paralelo para extraer información que le permita realizar tareas de vigilancia estratégica sobre Internet. Básicamente lo que hacen, es plantear el problema de búsqueda de información como un problema de optimización, el cual pueden resolver con técnicas evolutivas.
  En [35], proponen la utilización de técnicas evolutivas para apoyar el proceso de clasificación de documentos realizado por un sistema inmunológico artificial, específicamente, para evolucionar el sistema inmunológico artificial visto como el conjunto de detectores generados.

2.2.2.2. Análisis No Supervisado

A diferencia del análisis supervisado, las técnicas basadas en análisis no supervisado no requieren de un conjunto de patrones de salida para reducir el error del entrenamiento. Las técnicas no supervisadas son utilizadas principalmente para encontrar grupos de elementos no marcados: no existe una clasificación final contra la que se pueda hacer una comparación. Algunas de las técnicas basadas en análisis no supervisado son resumidas a continuación.

- **Clasificación Básica de Bayes (Naive Bayes):** Esta es una de las técnicas de clasificación con métodos probabilísticos más simple que existe. En general el método bayesiano busca encontrar la probabilidad de que un cierto elemento pertenezca a una cierta clase.
  En general, se asume la independencia de cada una de las clases, lo que simplifica bastante los cálculos involucrados. La clasificación de documentos se realiza como se propone en [25]. Algunos problemas de esta estimación, es que se asume la independencia de cada una de las clases evaluadas y que si el tamaño del vocabulario es mucho mayor que el número de documentos, los documentos pequeños no son tenidos en cuenta.
  Una aplicación al tratamiento de correos electrónicos es mostrada por Sahami [38], en donde se busca filtrar correos no deseados utilizando una aproximación bayesiana. En este caso, también se busca una estructura y una información específica en los mensajes, de modo que se puede aplicar correctamente el método de clasificación.

- **Redes SOM:** Los mapas auto-organizativos (Self-Organizing Maps o SOM) funcionan como redes neuronales artificiales pero no utilizan un patrón de salida conocido para entrenarse (entrenamiento no supervisado), es decir, clasifican de forma automática los datos de entrada presentados durante el entrenamiento y tiene la capacidad de ajustar nuevos datos a los patrones encontrados [39]. La arquitectura de la red puede ser lineal o corresponder a una malla bi-dimensional en donde la asociación entre las neuronas está dada por el concepto de vecindario [40] el cual puede ser lineal o bi-dimensional.
  El algoritmo SOM realiza aprendizaje competitivo donde no solamente los pesos, los que se pueden interpretar como centroides del grupo [41], de la neurona ganadora son adaptados, también lo son los de las neuronas pertenecientes al grupo (vecindario). Esto quiere decir que neuronas que están cerca en la malla son capaces de responder a entradas similares.
  Los mapas SOM tienen únicamente dos capas [40], la de entrada y la de salida. La de entrada es que recibe los patrones $\mathbf{X} = \{x_1, x_2, ...x_n\}$, el número de neuronas en esta capa será entonces n . El número de neuronas en la capa de salida depende de la organización (uni- o bi-dimensional) y de las características de la salida. La medida de similitud en esta técnica está dada por la

distancia del patrón buscado con el centroide del grupo con el que es comparado. La distancia puede ser medida utilizando diferentes medidas, algunas de las más utilizadas son:

- Distancia Euclídea: Esta es la medida natural de distancia. Simplemente determina si una muestra pertenece a un grupo determinado.
- Distancia Normalizada: Funciona de la que la distancia Euclídea, pero tiene en cuenta la dispersión del grupo.
- Distancia Minkowsky: Esta medida de distancia tiene en cuenta las desviaciones de cada una de las características y toma la mayor de ellas.

La explicaciones detalladas de los algoritmos y de las fórmulas de actualización pueden ser consultadas en [39] y [40].Las tareas que pueden ser resueltas con este tipo de técnicas son aquellas que son planteadas como un problema de agrupamiento, en el que no se conocen las clases existentes. En el caso de categorización de textos, lo que se hace primero es convertir cada texto en un punto en un espacio n-dimensional y luego aplicar el método específico.

2.2.2.3. Otras técnicas

Las técnicas explicadas anteriormente tienen enfoques con inspiraciones bien definidas para el tratamiento de textos, sin embargo, existen otras aproximaciones, como la propuesta en [42], en la que se utiliza el grafo de documentos que se generan como un paisaje. Esto permite además de realizar una agrupación, ver otras características de los grupos encontrados, como por ejemplo, relaciones entre grupos.

Masterson [43] propone una forma de extraer información de documentos que están relacionados entre si a través de un hilo conductor. Un tipo especial de éstos documentos son los correos. La idea se basa en la técnica de extracción de información que busca las características recurrentes en un texto.

Actualmente se está difundiendo una técnica de análisis conceptual de documentos llamada Análisis Formal de Conceptos (FCA, por sus iniciales en inglés) desarrollado por Wille [44] en 1982 , que busca modelar de forma matemática los conceptos contenidos en textos y presentar la relación de sus conceptos en forma de un grafo, le cual es generado a partir de ciertos atributos; un explicación muy interesante de los conceptos involucrados puede encontrarse en [44, 45, 46].

En [47] se propone un sistema de clasificación de colecciones de correos electrónicos basado en FCA. En este trabajo se expone la dificultad del trabajo con correos electrónicos debido a la informalidad del lenguaje utilizado y la alta contextualización del mismo; sin embargo, algo que merece más atención que el clasificador en si mismo, es la forma en que se debe analizar el grafo para extraer la información correcta.

2.3. Gestión de Conocimiento

La definición de conocimiento ha sido un tema discutidos por los filósofos desde la antigüedad hasta nuestros días, por lo que se convierte en uno de los problemas más importantes de la filosofía occidental y en lo que centran los estudios modernos de filosofía.

El conocimiento, ha sido uno de los símbolos y medios de progreso de la humanidad, permitiendo avanzar desde la invención de la rueda, hasta los viajes espaciales

Aunque no existe una definición clara y única de lo que es, si se sabe que representa un activo importante en las empresas, el cual, como el resto de activos, debe ser gestionado de la mejor manera posible de tal forma que represente una ventaja frente a otras compañías. Es decir, este activo es el que logra que en una organización, sus recursos de información generen ventajas e ingresos [48].

Es importante resaltar, que el conocimiento no existe por si mismo, es decir sin personas. Adicionalmente, el conocimiento es relativo a cada individuo, convirtiéndolo en una interpretación de una realidad particular o colectiva.

La gestión del conocimiento, utiliza diferentes fuentes para identificar y extraer el conocimiento. La fuente más evidente, son los datos y la información que de ellos se deriva. Cuando se tiene un conjunto de datos corporativos, se le aplican una serio de procedimientos algorítmicos que permiten su transformación en información. Estos algoritmos, comienzan a volverse complejos en la medida en la que los datos son transformados.

Figura 2.3: Datos, Información y Conocimiento. Adaptada de [48]

En la Figura 2.3, se puede ver el aumento de la abstracción de los datos y la información, y como los métodos computacionales con los que se debe trabajar se vuelven más complejos con el aumento de la abstracción.

2.3.1. Tipos de conocimiento

Según [49], existen dos tipos de conocimiento: el tácito y el explicito. La principal característica del primero es la forma en que es generado, ya sea través de la experiencia, a través de narraciones o a través de la exteriorización de información latente. El segundo, se refiere a todo conocimiento que está codificado de alguna forma, ya sea en documentos, reportes, bases de datos, audio y video, entre otros.

Adicionalmente, cada uno de dichos tipos de conocimiento tiene otra característica particular, que se refiere a la forma de ser almacenado y extraído. El conocimiento tácito, es almacenado de alguna forma en el cerebro, para hacer parte del inconsciente de la persona, de modo, que recodificarlo o explicitarlo, se convierte en una tarea compleja. Por otro lado, el conocimiento explícito, al ser un recurso físico, es fácilmente almacenable, ya sea en medios digitales o escritos, y por l tanto, es también fácilmente recuperable.

Sin embargo, la proporción en la que cada uno aporta es muy diferente. En la Figura 2.4, se muestra que la cantidad de conocimiento tácito disponible puede llegar hasta el 95 % del total, además que la forma en la que mejor se transfiere es oralmente.

2.3.1.1. Conocimiento tácito

El conocimiento tácito es aquél que es generado a partir de la experiencia y de la interiorización particular de conceptos, es decir, aprendizaje. Una de las principales características de este conoci-

Figura 2.4: Composición y transmisión de conocimiento. Adaptada de [48]

miento es la dificultad que tiene para ser extraído y representado. Sin embargo, es este conocimiento el utilizado para crear el conocimiento explícito [48].

Este tipo de conocimiento representa un gran valor para las compañías, ya que en el se encuentra la memoria corporativa, buenas y malas prácticas, experiencias en proyectos y situaciones particulares, entre otros. Sin embargo, este está "almacenado" en la experiencia particular de cada empleado, de modo que no es una tarea sencilla extraerlo, codificarlo y almacenarlo. Una razón, tal vez no la más importante, pero si las más común, es porque las personas se sienten seguras con ese conocimiento, es decir, se sienten indispensables y con algún nivel de poder en su entorno. Otra razón, es que el conocimiento, al ser personal, es almacenado y estructurado de formas particulares en cada persona.

Un ejemplo sobre la complejidad y extracción del conocimiento consiste en la forma en que una persona utiliza en bicicleta. En general, las personas aprenden a utilizar una bicicleta bien a través de la práctica y basándose en recomendaciones generales de alguien que ya aprendió. Sin embargo, una vez que se aprende, es muy complicado codificar dicho conocimiento de modo que sea transmitido a otras personas en forma de manual o tutorial.

Este es uno de los principales problemas en la gestión de conocimiento, como capturar correctamente y gestionar el conocimiento tácito; por ejemplo en [50] proponen un formulario para capturar información de expertos y poder extraer (o parte de) su conocimiento tácito y con esto construir una serie de redes sociales de conocimiento. Dada la dificultad de trabajar con este tipo de conocimiento, en [51] se muestra una tabla con aquellos conceptos que pertenecen al conocimiento tácito pero que pueden ser explicitados de alguna forma y aquellos que no; estos son el conocimiento tácito articulable y el no articulable.

- **Articulable:** El conocimiento tácito articulable es todo aquel adquirido a través de la educación recibida por un individuo. Este tipo de conocimiento puede ser reproducible de una forma relativamente sencilla. Una de las formas más sencillas de transmitir este conocimiento es a través de la renovación del ciclo de enseñanza, es decir, cuando un individuo transfiere a otros su conocimiento adquirido a través de relatos o explicaciones.

- **No articulable:** El conocimiento no articulable es todo aquel conocimiento adquirido por la experiencia individual, que se recoge a través del tiempo por el cual se ejerce alguna actividad. La principal característica de este tipo de conocimiento es que no puede ser fácilmente transmitido a otros individuos, ya que una persona sabe como realizar alguna actividad, pero no sabe exactamente como la hace y por eso no está en capacidad de transmitir fácilmente su conocimiento.

2.3.1.2. Conocimiento explícito

El conocimiento explícito es todo aquel conocimiento que está disponible en documentos, bases de datos o cualquier otro tipo de repositorio físico, que además, describe algún procedimiento o técnica.

Esto quiere decir que puede ser aprendido a través de libros o de artículos y es fácilmente extraible y administrable.

2.3.2. Ciclo de conocimiento

Según Nonaka [49], el conocimiento dentro de una organización se genera a partir de cuatro fases:

- **Socialización:** se refiere a la comunicación de conocimiento tácito a tácito, y por lo general, tiene lugar en discusiones de personas y en reuniones de equipos de trabajo[48]. Esta transferencia es la que se lleva a cabo cuando se narran experiencias entre personas, frente a frente, y no se produce conocimiento explicito.
- **Exteriorización:** esta es la fase en la que el conocimiento tácito se convierte en conocimiento explícito. Se presenta cuando se articula el conocimiento tácito, lo cual, generalmente, se realiza a través de diálogos, en los que se construyen objetos de conocimiento (intermediadores), en torno a los cuales se construyen ideas.
- **Comunicación:** en esta fase, el conocimiento explícito se convierte en conocimiento explícito, generalmente, a través del uso de la tecnología [48]. Un ejemplo en este caso, es el envío a través de correo electrónico un reporte escrito.
- **Interiorización:** en esta fase del ciclo, se convierte el conocimiento explícito en tácito. Esto se logra cuando se toma conocimiento explícito, como un reporte o un manual técnico, y se interioriza, de modo que se puedan deducir nuevas ideas o realizar acciones constructivas [48] como resolver problemas en áreas particulares del conocimiento.

En la Figura 2.5 se puede ver un esquema de este ciclo. Este ciclo no tiene un punto de inicio determinado, es decir, puede comenzar en cualquier punto. Se debe tener en cuenta, la forma en que el conocimiento explícito es codificado y el tácito difundido. La codificación tratará la forma en que se almacena el conocimiento de modo que pueda ser utilizado con posterioridad. Por otro lado, la difusión, consiste en fomentar la comunicación entre individuos que componen la organización de modo que se pueda crear el conocimiento tácito colectivo.

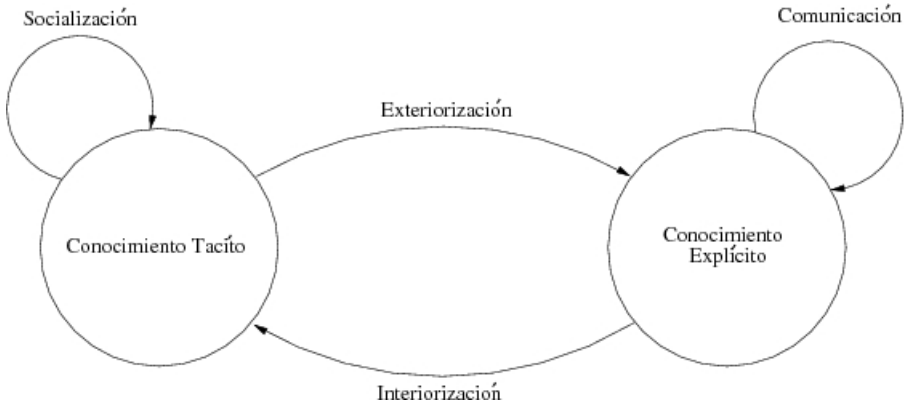


Figura 2.5: Ciclo de creación del conocimiento

[49]

Según Awad [48], se deben crear herramientas que permitan derivar conocimiento tácito del explícito. Es este uno de los principales objetivos de la gestión del conocimiento.

Capítulo 3

Arquitectura del Sistema de Extracción y Visualización de Información de Correos Electrónicos

> "Einstein's breakthrough was classic in that it sought to unify the elements of physical analysis, and it placed the older examples and principles within a broader framework. But it was revolutionary in that, ever afterward, we have thought differently about space and time, matter and energy" – Howard Gardner in Creating Minds, 1993.

Debido a la complejidad en el manejo de correos electrónicos y de la cantidad de información que ellos contienen, es necesario diseñar un sistema que permita extraer y visualizar la información contenida en un conjunto de correos electrónicos personales. Este sistema debe permitir extraer la información que permita construir, primero, la red contactos del dueño de los correos y segundo, identificar los temas en los que se encuentra divido el conjunto de mensajes de correo electrónico.

Adicionalmente, una vez se han extraído los dos tipos de información, es necesario que el sistema permita vincularlos, de tal forma que el usuario pueda ver, de forma intuitiva, las relaciones existentes entre la red de contactos y los temas del conjunto de mensajes.

En este capítulo se presenta la problemática relacionada al manejo de correos electrónicos personales, la definición de los requerimientos funcionales del sistema de extracción y visualización, la descripción de la arquitectura global del sistema y la descripción del proceso de extracción, transformación e indexado del conjunto de datos.

3.1. Requerimientos del Sistema

El sistema de extracción de información debe implementar un proceso de extracción que incluya fases de pre-procesamiento, estandarización, indexamiento e identificación de la información de redes de contactos y temas de los mensajes de correo electrónico. Estas fases generales permiten controlar el proceso y agregar cierta modularidad, especialmente a lo relacionado con la fuente de datos, es decir, el conjunto de datos.

Para plantear los requerimientos del sistema se hará una división por cada una de las fases mencionadas.

- **Pre-procesamiento y estandarización:** el sistema de pre-procesamiento debe permitir convertir a un formato estándar correos provenientes de diferentes fuentes. Es decir, que todos los conjuntos de mensajes utilizados sean convertidos, a través de una interfaz, a un solo formato, de tal forma que todos los pasos siguientes del proceso de extracción, identificación de información

y visualización, sean totalmente independientes de la fuente primaria de información.
El formato seleccionado para esto es XML, o una extensión del mismo, de modo que resulte sencillo, en términos computacionales, que un computador, por medio de un algoritmo dispuesto para ello, pueda leer los correos y extraer la información requerida.

- **Indexamiento:** el sistema debe permitir que los mensajes de correo electrónico que han sido convertidos a un formato estándar puedan ser recuperados a través de procedimientos genéricos durante la ejecución del proceso de extracción. Este indexamiento debe permitir acceder a cualquiera de los mensajes del conjunto con una llave única y adicionalmente, acceder a cualquiera de los campos contenidos dentro del mensaje de correo.

- **Identificación de información:** el sistema debe permitir identificar la información necesaria para construir las redes sociales y los mapas utilizados para describir los temas contenidos en los mensajes de correo electrónico. En el caso de las redes de contactos, es necesario que la información extraída permita representar un grafo en el que los nodos sean personas y los arcos representen las relaciones entre las personas. En el caso de los temas, se requiere que se pueda disponer de la información del asunto y del cuerpo del mensaje, de tal forma que se pueda realizar un agrupación de mensajes y una posterior identificación de temas.

Cabe resaltar que todos los correos electrónicos poseen cierta cantidad de conocimiento tácito, el cual, está en gran medida representado por las relaciones que se generan entre ellos y por los objetos por los que dichas relaciones se dan. Este conocimiento, al no estar codificado, no es fácil de extraer.

Para extraer dicho conocimiento, se debe desarrollar un sistema de análisis de correos electrónicos que permita extraer las características relevantes, como información de contactos y temas, entre otros, de los mensajes y de los documentos adjuntos de modo que se puedan asociar con las relaciones identificadas en las redes sociales extraídas de los mensajes de correo electrónico. Con esta asociación se busca darle un sentido más amplio a las relaciones identificadas entre personas, permitiendo que el usuario pueda resolver preguntas del tipo ¿Quién sabe qué?.

3.2. Arquitectura Global de extracción de información

Un mensaje de correo electrónico contiene información acerca del remitente y los destinatarios, tal como el nombre y la dirección electrónica, el asunto, el cual está ligado a la intención con la que el mensaje es enviado, y el mensaje mismo, en donde se incluye la información que se quiere transmitir a la otra u otras personas. Adicionalmente, incluye campos como la fecha en el que fue enviado, el tipo de servidor de correo a través del que fue enviado al que llegó, la ruta de retorno y otros detalles técnicos. En el caso particular de este trabajo, solo se tendrá en cuenta la información relacionada con las personas, el asunto y el cuerpo del mensaje. Entonces, un mensaje m estará definido por la tupla:

$$m = \langle R, D^+, A, C \rangle \tag{3.1}$$

donde R, corresponde al remitente del mensaje, D^+, representa la lista no vacía de destinatarios, A, es el asunto del mensaje y C, representa el cuerpo del mensaje.

Hasta ahora, se ha definido solamente la composición de un mensaje de correo electrónico, sin embargo, el objetivo de este trabajo es mostrar el conocimiento y la información no evidente contenida en un conjunto de mensajes de correos electrónicos, los cuales, pertenecen a una sola persona. Estos son entonces, mensajes que una persona, que es la dueña de dichos mensajes, ha recibido y/o ha enviado.

Entonces, un conjunto M de mensajes, puede ser definido por:

$$M = \bigcup_{i=1}^{n} m_i \tag{3.2}$$

donde n es el número total de mensajes, es decir, $\|M\| = n$.

Sobre dicho conjunto M, se realizan todas las operaciones de transformación y extracción requeridos para obtener la red de contactos y una categorización de los temas contenidos en ellos.

La red de contactos corresponde a la información acerca de las personas contenida en los mensajes de correo. Estas personas son incluidas en los mensajes al ser destinatarios y/o remitentes de mensajes. Utilizando dicha información es posible identificar las relaciones entre personas, las cuales, se dan cuando las personas aparecen en un mensaje y, al menos una de ellas es el remitente y el resto, pertenece al conjunto no vació de destinatarios.

Con el conjunto de personas y sus relaciones es posible definir un grafo dirigido $G\langle V, E\rangle$, en donde V corresponde al conjunto de de vértices, los cuales representan a las personas en los mensajes, y E, el conjunto de arcos, que representa el conjunto de relaciones entre personas.

Cada uno de los arcos tiene un peso dado por la frecuencia de envío de mensajes de una persona a, a una persona b.

La categorización de los temas busca identificar los diferentes tópicos que son tratados por la persona dueña con otras personas. Estos temas pueden ser bastante variados, sin embargo, pueden estar relacionados unos con otros de diferentes formas y en medidas diferentes.

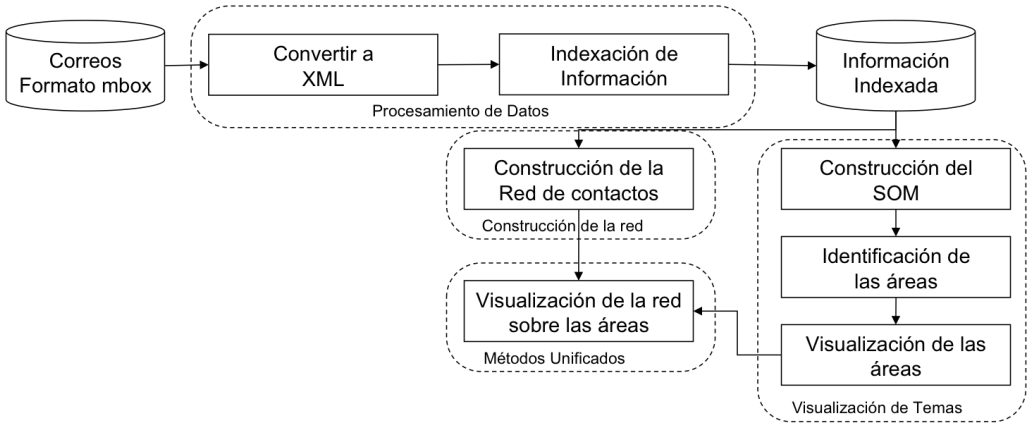


Figura 3.1: Arquitectura funcional del sistema

En la Figura 3.1, se muestra la arquitectura funcional del sistema propuesto y cada uno de sus componentes principales. Este sistema abarca desde el pre-procesamiento de los mensajes de correo electrónico, es decir, su estructuración, hasta la unificación de los esquemas de visualización de las redes de contactos y los temas en los mensajes, pasando por la extracción de la información y el conocimiento contenido en los mensajes de correo electrónico.

En las siguientes secciones se describe cada uno de los componentes del sistema mostrado en la Figura 3.1.

3.3. Proceso de Extracción, Transformación e Indexación de la información de los Mensajes

El conjunto de mensajes de correo puede estar en un formato particular, en el cual, se definen y ordenan los diferentes campos de cada mensaje de una forma diferente. Estas variaciones en los formatos hace que sea necesario definir un formato único para almacenar los mensajes y que permita que el procedimiento de extracción de información sea el mismo para todos los mensajes.

En este caso, el origen de los mensajes es un conjunto de correos electrónicos definidos en formato RFC 822, el cual es un formato general para almacenar correos en un servidor POP3 y dentro de los campos definidos en este formato se encuentran:

From: Contiene la información del remitente del correo. Este campo es obligatorio y no puede ser vacío.

To: Este campo contiene una lista no vacía de direcciones de correo electrónico los cuales son los destinatarios principales. Se supone que las personas que son incluidas en este campo tienen cierta obligación de contestar el mensaje que les envían, o al menos, saben que el mensaje es específicamente para ellos.

Cc: Este es un campo utilizado para enviar correos a personas que deben ser informadas de determinada situación o que requieren estar al tanto de cierta información. Este campo es opcional dentro de la construcción del mensaje.

BCc: En este campo se incluyen las personas a las que se les quiere hacer llegar cierta información pero sin que los demás destinatarios lo sepan.

Subject: En este campo se especifica el asunto del mensaje, el cual, es utilizado para informar a los destinatarios la razón principal del envío.

Body: Este campo contiene la información principal del mensaje. En este está el mensaje como tal y es la razón real y final para el envío del mensaje.

Date: Este campo contiene la fecha en la que el mensaje fue enviado.

En el formato original, se encuentran en texto plano y para obtener cada campo es necesario recorrer línea por línea el mensaje. Esto en mensajes muy grandes puede ser complejo.

A partir de la información de cada mensaje se puede hacer una conversión a un formato estándar, que además, tenga una estructura única, y así, sea posible acceder rápidamente a cualquiera de los campos que se requieran durante los procesos de extracción de información.

El formato más adecuado para esto es XML, que además de ser estructurado, permite acceder a cualquier campo del mensaje de una forma eficiente. Un ejemplo de un mensaje transformado en XML se puede ver en el listado siguiente:

```
<message>
    <date>
        Enero 1 de 2007
    </date>
    <from>
        remitente@server.com
    </from>
    <to>
        destinatario@otrosrv.com
    </to>
    <subject>
        Hola
    </subject>
    <body>
        Este es el cuerpo del mensaje...
    </body>
</message>
```

Para extraer la información de un mensaje particular, se accede al campo requerido, por ejemplo <*body*>, en cuyo caso se obtendrá el cuerpo completo del mensaje y ya es posible utilizar esta información para procesarla y ejecutar los diferentes procedimientos que se definen posteriormente para unificar la información.

De esta forma, al menos desde el punto de vista computacional, es más sencillo cargar cada uno de los elementos del mensaje., de otra forma, sería necesario recorrer cada mensaje y registrar el número de bytes leídos cuando se encuentra un nuevo campo, lo que resulta demorado y poco eficiente.

3.4. Construcción y Visualización de Redes de Contactos

Una vez obtenida la información necesaria para construir la red de contactos, es posible construir la red de contactos de cada uno de los mensajes y del conjunto de correos. La operación de construcción de la red de contactos se muestra con mayor detalle en el Capítulo 4.

Para construir la red de contactos se toman los datos de origen y destino de los correos electrónicos y se construye una matriz de adyacencia en la que se especifican los orígenes y los destinos de las conexiones entre los nodos. Note que en este caso, las comunicaciones entre personas no son unidireccionales, razón por la cual la matriz de adyacencia no es simétrica.

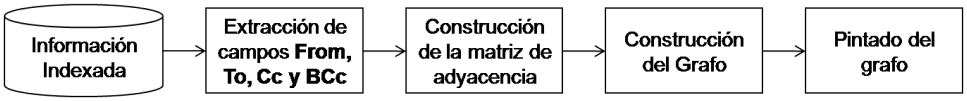


Figura 3.2: Diagrama de bloques del proceso de construcción y visualización de la red de contactos

En la Figura 3.2 se muestra el diagrama general de bloques del proceso de construcción y visualización de las redes de contactos. Debido a que la matriz de adyacencia puede ser muy grande, se utiliza un esquema de representación no matricial en donde cada cada nodo del grafo tiene asignada una lista de nodos a los cuales se conecta. Esta representación permite además asignar diferentes características a cada nodo en la red, lo que será útil más adelante cuando se integren los esquemas de visualización de redes y temas.

```
<nombre de la red>;<lista de nombres de características>
<nodo origen>;<lista de nodos destino>;<lista de características*>
```

Figura 3.3: Definición del archivo de representación de la matriz de adyacencia

En la Figura 3.3 se muestra el esquema utilizado para definir la matriz de adyacencia. En la primera línea se describe el archivo y se especifican los identificadores de las diferentes características asignadas a los nodos.

Una vez construida la matriz de adyacencia, se procede a desplegar el grafo descrito por ella utilizando alguno de los algoritmos diseñados para ello, los cuales serán discutidos en el Capítulo 4, inclusive, se puede utilizar un algoritmo aleatorio, en el a cada nodo de la red se le asigna una posición(x, y)arbitraria dentro del espacio de pintado.

3.5. Identificación y Visualización de Temas

Cada uno de los mensajes de correo electrónico ha sido enviado con una finalidad, sin importar si es o no solicitado (SPAM), y por tal razón puede pertenecer a una categoría temática. Al unir

todos los mensajes de correo se pueden agrupar por dichas categorías temáticas, las cuales deben ser identificadas.

En la Figura 3.4 se muestra el esquema general del proceso de identificación y visualización de las áreas temáticas de los mensajes. El primer paso es la transformación de los mensajes a un formato que pueda ser utilizado por los algoritmos de entrenamiento. En este caso, el cuerpo de los mensajes son convertidos en vectores para luego ser agrupados a través de un SOM.

Luego, utilizando la información de grupos arrojada por el SOM, se identifican y definen las áreas de temas, incluyendo la identificación de los tamaños y sus centros de masa respectivos.

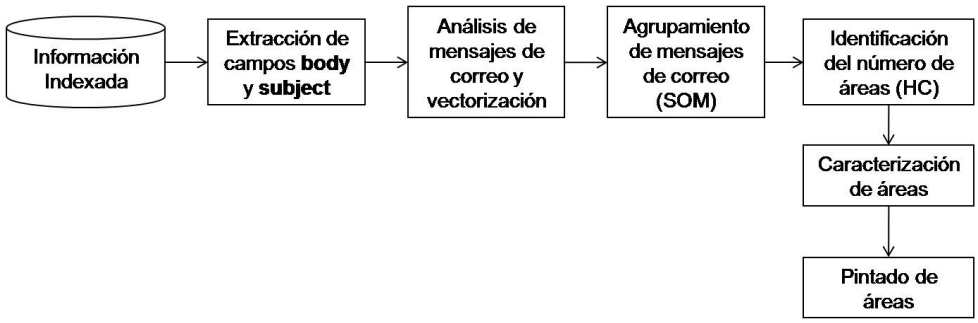


Figura 3.4: Diagrama de bloques del proceso de identificación y visualización de los temas de los correos electrónicos

Una vez identificadas las categorías se deben visualizar de tal forma que las que guarden cierta similitud, estén más cerca unas de otras, y las que no, estén alejadas. Para esto, cada una de las categorías se define como un área geométrica, la cual tendrá un centro de masa, utilizado posteriormente para la construcción visual, y un tamaño determinado por el número de correos que pertenecen a la categoría representada por el área.

El proceso de construcción y visualización de las áreas será descrito con detalle en el Capítulo 5.

3.6. Unificación de Esquemas de Visualización

Una vez se han creado los esquemas de visualización de redes de contactos y de áreas de temas, se integran un solo espacio de visualización, de tal forma que se puedan relacionar tanto la red de contactos, con sus vínculos entre personas, y los temas de los que las personas en dicha red intercambian mensajes.

A través de la integración de la red de contactos y de los mapas de temas, en los que ese encuentran las áreas temáticas, es posible para los usuarios asociar personas dentro de su red a un conjunto de temas particulares y saber los temas de los que sus diferentes contactos hablan en su red. Así mismo, su posición en el mapa le indica cual es tema del que más habla siendo el dueño de los mensajes.

El proceso de unificación de los esquemas de pintado es detallado en el Capítulo 5.

Capítulo 4

Construcción y Visualización de la Red de Contactos

> "Knowledge resides in the user and not in the collection [of information]. It is how the user reacts to a collection of information that matters." – Churchman, C.W. (1971). The Design of INQUIRING SYSTEMS: Basic Concepts of Systems and Organization, Basic Books, New York, NY, p. 10.

El primer paso para crear un esquema unificado de visualización corresponde a la construcción de la red de contactos y posteriormente definir un esquema para su visualización. La construcción de contactos se hace a partir de los campos que describen a las personas relacionadas con el mensaje, es decir, el remitente y los destinatarios.

Esta extracción se hace para cada uno de los mensajes, de tal forma que para cada uno de dichos mensajes se construye el grafo que lo representa y posteriormente, se hace una operación de unión sobre todos los grafos y se construye el grafo que representa todo el conjunto de mensajes.

Una vez definido el grafo de procede a ejecutar diferentes algoritmos de pintado, buscando que la interpretación de la red de contactos sea más sencilla para los usuarios. De esta forma se prueban algunos algoritmos existentes, se proponen otros y se plantea un esquema de coloreado para los nodos basado en la distancia del nodo central.

4.1. Construcción de la Red de Contactos

Los contactos en un conjunto de mensajes de correos electrónicos, corresponden a todas las personas y/o organizaciones, cada una de ellas representadas por una dirección de correo electrónico, dentro del conjunto de mensajes. De esta forma, cada persona, ya sea remitente o destinatario, hará parte del conjunto de contactos.

Según esto, el conjunto de contactos P está definido por:

$$P = \{p_i \mid p_i \, es \, direccion \, de \, correo \, \forall i \geq 0\} \tag{4.1}$$

donde p_i es el contacto $i-$ésimo. A partir de esto se puede establecer que $P \subseteq V$, donde V es el conjunto de vértices del grafo dirigido que representa la red de contactos.

Para extraer los contactos de un conjunto de mensajes, es necesario definir inicialmente los campos que permiten identificar a uno de dichos contactos en un mensaje particular. Según el RFC 822, cada uno de los mensajes en formato mbox deben contener un remitente, el cual, está determinado por la palabra *From* dentro del mensaje y un conjunto no vacío de destinatarios, el cual, está dividido en tres subconjuntos: los destinatarios principales, los cuales está determinados por la palabra *To*, los destinatarios a los que se les envía una copia del mensaje, determinados por la palabra *Cc,* y los

destinatarios a los que se les envía una copia oculta del mensaje, los cuales están determinados por la palabra *BCc.*

Cabe anotar, que aunque la lista de destinatarios no puede ser vacía, el único de los subconjuntos descritos anteriormente que no puede ser vacío, es el de los destinatarios principales.

Para extraer los contactos, se supone que ya han sido procesos y convertidos a un formato estándar, el cual, permite acceder de forma rápida a cada uno de ellos y obtener la información que se requiera. La operación de extracción se realiza para cada uno de los correos, de modo que cada uno de los mensajes tiene un conjunto de contactos, o personas, que están relacionadas entre si, así, lo que se va a tener al final, es la unión de todas las sub-redes de contactos encontradas en cada mensaje. Esto quiere decir,que cada uno de los mensajes que pertenecen a un conjunto de mensajes particular, contiene un red de contactos, la cual, es conservada dentro de la red de contactos que representan al todo el conjunto.

Por lo tanto, la red de contactos está determinada por:

$$R = \bigcup_{M} r_{m_i} \tag{4.2}$$

donde, M es el conjunto de mensajes, definido por la Ecuación 3.2 y r_{m_i} es la red de contactos del mensaje $i-$ésimo.

Cada uno de los campos tiene una interpretación particular desde el punto de vista del análisis de redes sociales. Debido a que en los mensajes de correo electrónico solo existe un remitente y uno o más destinatarios, cada uno de los mensajes forma una red social llamada egocéntrica, lo que quiere decir que existe un solo nodo que concentra la mayor proporción de vértices salientes [1]; este nodo, en términos del análisis de redes sociales es conocido como *ego*. Los otros nodos, es decir, aquellos a los que les es enviado el mensaje, son conocidos como *alter*.

Algo interesante resulta que al aplicar la ecuación 4.2, la red que representa el conjunto total de mensajes, sigue siendo egocéntrica, y el nodo ego corresponde al dueño de la cuenta de correo que se está analizado.

Dado que las redes de contactos están basadas en grafos, es posible efectuar ciertas operaciones sobre ellos que permiten establecer ciertos comportamientos e indicadores con el fin de facilitar el análisis de dicha red. Las operaciones realizadas, buscan identificar aquellos nodos que tienen una cierta importancia dentro de la red, es decir, aquellos que me permiten alcanzar a otros nodos, los que concentran el mayor número de comunicaciones o los que permiten agrupar a otros nodos de la red, entre otras.

Adicionalmente, es posible realizar operaciones que permitan determinar sub-grupos de nodos, los cuales, tienen determinada cercanía por la frecuencia de comunicaciones o fuerza del arco entre ellos. Dichas operaciones se resumieron en la Sección 2.1.3.

Computacionalmente hablando, la forma más sencilla de representar una red social es a través de una matriz de adyacencia, en la cual, cada componente i, j contiene un valor binario $(1\, o\, 0)$ que indica la presencia o la ausencia respectivamente de una conexión entre el nodo i y el nodo j.

Esta representación, aunque es sencilla, resulta costosa en términos de uso de espacio, y dado que es un grafo dirigido, no se puede asegurar que $\mathbf{M} = \mathbf{M}^T$, no es posible utilizar únicamente la información que está arriba de la diagonal superior, por lo que se puede utilizar un esquema de lista, en donde todos los nodos tienen una lista de nodos asociados. Esta representación reduce el espacio a utilizar, aunque hace que las operaciones sobre el grafo sean complejas.

4.2. Pintado de la Red de Contactos

Se han desarrollado diversos algoritmos para organizar y desplegar grafos, los cuales pueden ser utilizados en grafos de cualquier tipo. En este caso se tiene grafos que representan redes sociales egocéntricas, por lo que se buscan algoritmos que conserven la estructura de la red y permitan verla de una forma organizada.

A continuación se presentan algunos de dichos algoritmos. En primer lugar, dos basados en fuerza: el Eades y el de Fruchterman & Reingold y luego, escalamiento multidimensional, basado en las distancias (similitudes en algunos casos) de los nodos del grafo.

Para las pruebas realizadas de con los algoritmos mostrados en esta sección se ha utilizado un sub-conjunto de datos reducido tomado del conjunto de datos de ENRON[1] en cual es un conjunto de correos electrónicos que la comisión americana de energía decidió liberar luego de su investigación. Este conjunto de datos está compuesto por los correos de 150 usuarios, que suman cerca de 500000 mensajes, todos en el formato *mbox*.

4.2.1. Algoritmos Convencionales de Visualización de Redes Sociales

En esta sección se presentan otros algoritmos de visualización de redes sociales que han sido utilizados en diversas aplicaciones. Aunque fueron desarrollados con otros fines, vale la pena incluirlos dentro de la experimentación de tal forma que se tengan puntos de comparación con los algoritmos desarrollados para asociar los temas de un conjunto de mensajes de correo electrónico y la red de contactos contenida en ella.

En primer lugar se presenta el algoritmo de Eades, el cual utiliza una aproximación basada en el principio de Hooke. Luego, se presenta el algoritmo de Fruchterman & Reingold, el cual busca utilizar de la mejora manera posible el espacio disponible de pintado, y finalmente, un algoritmo basado en Multi-Dimensional Scaling, el cual, utiliza las similitudes entre nodos para representar la estructura de la red.

4.2.1.1. Algoritmo de Eades

Uno de los primeros algoritmos desarrollados, y ampliamente utilizado, es el propuesto por Eades [52], en el que trata cada uno de los arcos como un resorte de acero, buscando aplicar el principio de Hooke. Con esto, convierte el grafo en un sistema mecánico que tiende hacia niveles de baja energía y cuando llega al mínimo nivel, se detiene.

La gran desventaja de este algoritmo es que calcula la energía total del sistema, por lo que su complejidad es de $\Theta(V)$ para los vértices y $\Theta\left(N^2\right)$ para los nodos, de tal forma que el tiempo de ejecución para redes sociales grandes puede ser alto.

En la Figura 4.1 se puede ver un ejemplo del algoritmo de Eades. En este algoritmo, cada una de las sub redes tiene una masa propia, por lo cual, a cada una de ellas se aplica el algoritmo como si fueran un elemento atómico, lo que hace que tengan un efecto repulsivo al igual que entre los nodos, haciendo que se desplacen en la forma que las flechas lo indican.

4.2.1.2. Algoritmo de Fruchterman & Reingold

El algoritmo de Fruchterman & Reingold [2], no asegura que se reduzca el número de cruces entre los arcos del grafo, sin embargo, logra una buena distribución de los nodos dentro del marco del dibujo. Su complejidad está dada por $\Theta(N \log N)$.

Este algoritmo calcula las distancias entre nodos utilizando la siguiente ecuación:

$$C\sqrt{\frac{area}{N}}$$

donde C es una constante hallada experimentalmente y N es el número de nodos.

Por esta razón, se asegura que se utilice de la mejor manera el área de pintado, aunque no conserva la estructura de la red dada por la matriz de similitud si es que esta existe.

En la Figura 4.2 se muestra el resultado del algoritmo de Fruchterman & Reingold. Dicha organización no representa necesariamente la estructura real de la red social en términos de distancias

[1] Disponible en: http://www.cs.cmu.edu/~enron/

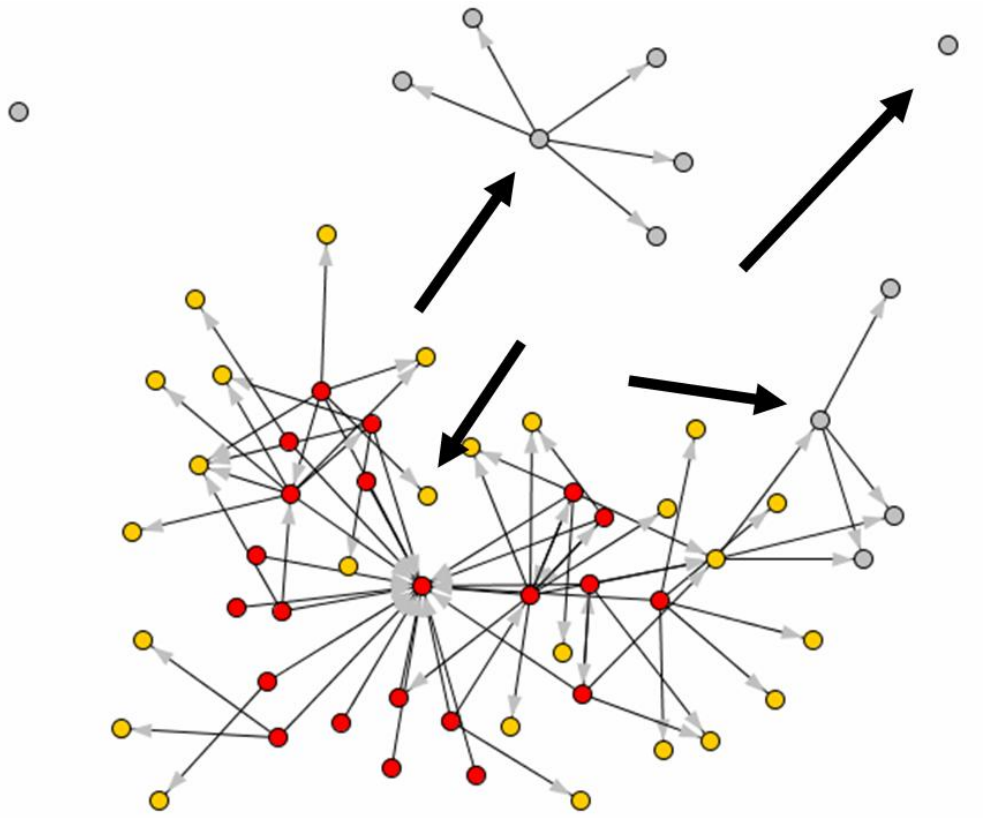


Figura 4.1: Comportamiento de los nodos en el algoritmo de EADES. Laas flechas indican la dirección de desplazamiento de las sub redes.

y posición geográfica, sin embargo, utiliza muy bien el área de pintado y hace que el grafo quede claramente pintado.

4.2.1.3. Escalamiento Multidimensional

Existen otro métodos de visualización en los que importa más conserva la estructura original de la red. Es el caso de dos de ellos, estudiados por Freeman [53]. El primero es llamado *Singular Value Decomposition* – SVD y el otro, escalamiento multidimensional (MDS por sus iniciales en inglés).

Ambos algoritmos tienen como entrada una matriz de $n \times n$, donde n es el número de nodos en el grafo, con las similitudes y/o distancias entre cada uno de ellos. Se busca entonces generar n puntos $m-$dimensionales tales que dichos puntos representen de la mejor manera posible dicha matriz.

El método MDS fue propuesto por Cox [3], se busca, dada una matriz de $n \times n$ de similitudes o disimilitudes, un conjunto de puntos en un espacio $m-$dimensional. En este caso, se busca llevar la matriz de disimilitudes δ a un conjunto de puntos p, cuyas distancias d reflejen de la mejor forma dichas disimilitudes.

Este proceso puede ser llevado a cabo a través de dos métodos. El primero, conocido también como métrico, es realizado a través de operaciones entre matrices. El inconveniente con este método es que el resultado no es exacto, es decir, los puntos encontrados no reflejan la distancia de la matriz de disimilitudes. El segundo, conocido como no métrico, utiliza algún método de optimización en gradiente descendente, por ejemplo *simulated annealing*. Este método necesita como entrada, un conjunto inicial de puntos, a partir de los cuales se van a buscar los que más se acerquen a la matriz de disimilitudes.

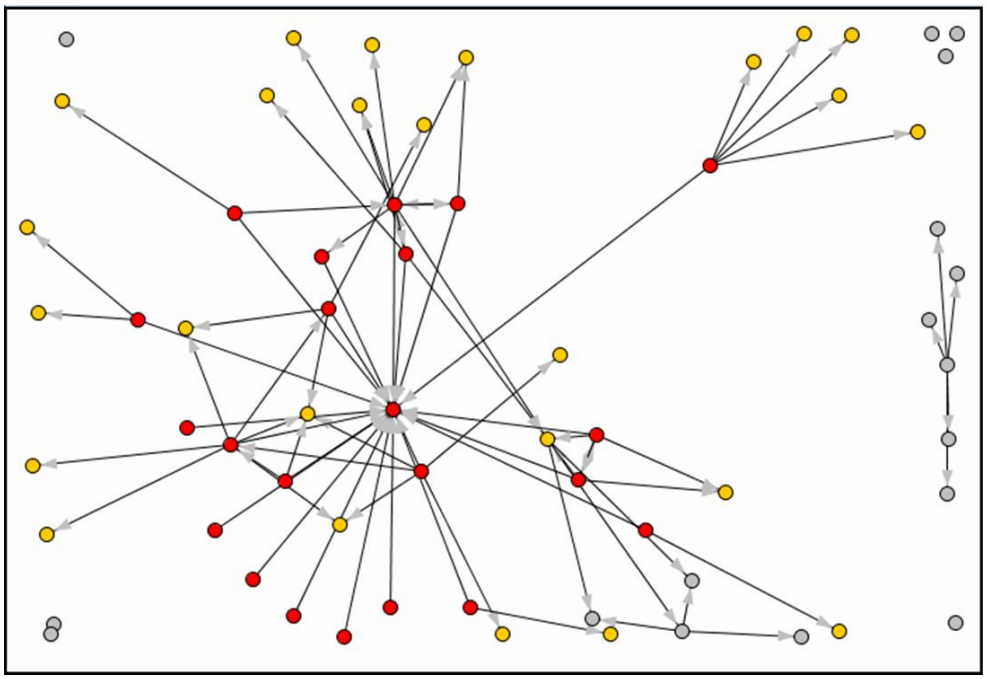


Figura 4.2: Organización basada en el algoritmo de Fruchterman & Reingold

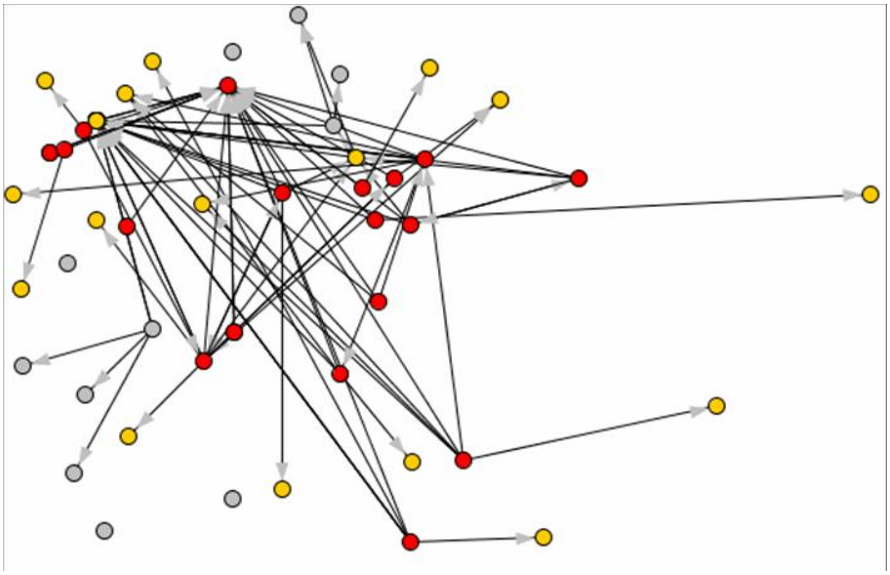


Figura 4.3: Organización basada en el algoritmo MDS

En la Figura 4.3 se muestra el resultado del algoritmo de organización utilizando MDS. En este caso se utiliza tanto el métrico como el no métrico.

4.2.2. Propuestas para el Manejo Gráfico de Redes Sociales

Una vez presentados algunos de los algoritmos convencionales para pintado de grafos, y particularmente, de redes sociales, se presentan otros métodos y esquemas de manejo gráfico que permitan identificar información relevante en términos del presente trabajo.

Primero se presenta un esquema de coloreado de los nodos que permite identificar el nodo que representa el dueño de los correos y los nodos de las personas que intercambian información con él. Luego se presenta un algoritmo que busca organizar los nodos dentro de anillos concéntricos según el grado de conexión con el dueño de los mensajes de correo. Finalmente, se presenta un algoritmo que permite ubicar una red social en un campo de fuerzas, el cual, está dividido en áreas geométricas, y cada uno de sus centros de masa actúan como puntos de atracción de los nodos de la red.

4.2.2.1. Coloreado de los nodos

Para facilitar la visualización de la red de contactos, es una buena práctica colorear los nodos de la red siguiendo algún patrón o alguna medida propia de la red, como el grado. En este caso, se utilizará como patrón el grado de separación de los alter (los nodos diferentes al nodo que representa al dueño de los mensajes de correo) y el ego (dueño de los mensajes de correo).

La idea es asignarle al ego un color particular, al los alter que se encuentran a un grado de separación del ego otro color, en este caso el rojo, a los alter que se encuentran a dos grados de separación serán de color amarillo y así sucesivamente. Aunque se pueden definir varios grados de separación, dado que la red contactos generada por un conjunto de mensajes de correos electrónicos es egocéntrica, no existan más de cuatro grados de separación.

Para determinar el color de un nodo particular se utiliza la matriz de adyacencia del grafo dirigido de la red de contactos. Esta es una matriz binaria no simétrica (o al menos en general no lo es) con la que se puede representar la totalidad de los nodos y las relaciones entre ellos. Una matriz de adyacencia puede ser como:

$$\begin{array}{c|ccccc} & e & a_1 & a_2 & \cdots & a_n \\ e & \mathbf{0} & 1 & 1 & \cdots & 0 \\ a_1 & 0 & \mathbf{0} & 0 & \cdots & 1 \\ a_2 & 1 & 0 & \mathbf{0} & \cdots & 1 \\ \vdots & \vdots & \vdots & \vdots & \ddots & 1 \\ a_n & 1 & 1 & 1 & \cdots & \mathbf{0} \end{array}$$

donde e corresponde al nodo ego y los $a_k, k = 1, 2, 3, ..., n$ corresponden a los alter. Entonces, a cada uno de los nodos del grafo se le puede asignar un índice K el cual puede variar de 0 a n, donde n es el número de nodos alter y el índice a_0 será utilizado para el nodo ego.

En primer lugar se va a definir la función $Ind\,(N)$, que retorna el conjunto de sub-índices del conjunto de nodos N. Por ejemplo, sea $N = \{a_2, a_4, a_7\}$, entonces la función $Ind\,(N)$ es igual a:

$$Ind\,(N) = \{2, 4, 7\}$$

Sea G_0 el conjunto de nodos que se encuentran a 0 grados de separación, es decir, el nodo *ego*, los nodos que se encuentran a un grado de separación G_1 estarán dados por:

$$G_1 = \bigcup_{j=1}^{n} \{a_j : M_{j,0} = 1 \vee M_{0,j} = 1\} \tag{4.3}$$

donde$M$ es la matriz de adyacencia. Esto quiere decir que solo los nodos que tengan conexiones, ya sean entrantes o salientes, con el nodo ego estarán en este conjunto.

Los nodos que tienen grado dos de separación serán aquellos diferentes a los que pertenecen al conjunto definido por la ecuación 4.3 y que tienen algún vínculo con con los nodos que se encuentran a un grado de separación.

Los nodos que tienen grado dos de separación están dados por:

$$G_2 = \left(\bigcup_{k \in Ind(G_1), j \in K} \{a_j : M_{j,k} = 1 \vee M_{k,j} = 1\} \right) - G_1 \tag{4.4}$$

Esto quiere decir que el conjunto de nodos de grado dos de separación será la unión de todos los nodos que tengan relación de un grado de separación con los nodos de grado 1 excluyendo los nodos que ya están en el conjunto G_1. De esta forma se puede definir una función recurrente para calcular el color de todos los nodos del grafo:

$$G_i = \left(\bigcup_{k \in Ind(G_{i-1}), j \in K} \{a_j : M_{j,k} = 1 \vee M_{k,j} = 1\} \right) - \bigcup_{l=1}^{i-1} G_l \tag{4.5}$$

Con la ecuación 4.5 se realiza el pintado de los nodos de la red. Esto permite dividir los nodos en categorías diferentes, de tal forma que se puede establecer la estructura de la red de contactos de una forma sencilla, aún cuando se tienen redes de contactos de gran tamaño.

4.2.2.2. Algoritmo de fuerza basado en anillos

Dado que en este caso, las redes sociales son egocéntricas, tienen un nodo central, que corresponde al individuo dueño de los mensajes de correo electrónico. Este individuo posee el mayor grado de toda la red ya que es el que más envía y el que más recibe mensajes. Por esta razón, se ha diseñado una forma diferente de organizar el grafo.

El primer paso, consiste en ubicar el nodo de mayor grado en el centro del área de pintado, y seguidamente, ubicar el resto de nodos teniendo en cuenta el grado de vecindad con el nodo central.

El grado de vecindad se refiere al número de saltos que es necesario dar desde o hacia el nodo central para llegar a o desde otro nodo. Por ejemplo, un individuo que está a un salto, tendrá grado de vecindad 1.

Con la definición de los grados, se puede plantear la creación de anillos de fuerza, que van a actuar como límites en los cuales cada nodo, según su grado, podrá estar. Así, el nodo central, estará solo en el anillo cero, en el siguiente anillo estarán los vecinos de grado uno y así sucesivamente.

El siguiente paso consiste en definir la distancia entre nodos conectados. Para esto se tiene en cuenta el grado de vecindad y el radio del primer anillo. Se supone que los anillos son círculos concéntricos de radio r.

La distancia entre el nodo $i-$ésimo de grado j con el nodo central, $d_{i,j}$ estará dada por

$$d_{i,j} = \beta_j r \left[(j-1) + \alpha_i \left(j+1 \right) \right] \tag{4.6}$$

donde α_i corresponde a un valor que depende del número de mensajes intercambiados, teniendo en cuenta el origen de cada mensaje, y la frecuencia de dichos intercambios en un intervalo de tiempo y $\beta_j > 0$ corresponde a un valor que depende de la densidad de elementos de grado j, permitiendo aumentar o reducir el radio de un determinado círculo según sea la cantidad de nodos de dicho grado.

Como el enfoque del estudio son los correos electrónicos personales, se eligieron de forma aleatoria dos de dichos usuarios. Con ambos conjuntos se realizaron las pruebas de estandarización, para luego realizar pruebas de funcionamiento de los algoritmos mencionados.

Las primeras pruebas realizadas fueron utilizando el algoritmo basado en fuerza. Para esto se utilizó uno de los conjuntos de datos reducidos, únicamente la carpeta de la bandeja de entrada. Esto con el fin de observar el comportamiento de los algoritmos pero teniendo en cuenta su complejidad computacional.

En la Figura 4.4 se muestra el resultado del algoritmo de organización con anillos de fuerza para el conjunto de datos reducido. Los nodos de color rojo, corresponden a los individuos que tiene grado de vecindad uno, los amarillos, los que tienen grado de vecindad dos.

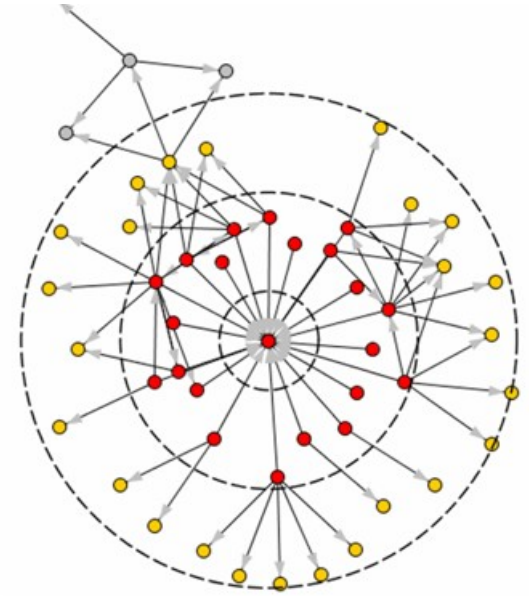


Figura 4.4: Ejemplo del algoritmo de fuerza basado en anillos

Se puede ver como todos los nodos están fuera del anillo cero, que contiene el central, y cada grupo ocupa su propio anillo. Esto mejora la visibilidad en una red de este estilo.

Esta versión del algoritmo utiliza un valor de $\beta = 1$, de modo que el radio crece de manera constante con respecto al grado, sin importar la densidad de los individuos de dicho grado.

Cabe señalar, que las vecindades, o saltos, representadas por los colores no se refieren a rutas directas de comunicación entre las personas representadas por al red sino que son sirven para indicar la separación entre dos personas, es decir, podría no existir una ruta directa desde el nodo que representa al dueño de los mensajes y uno de los nodos amarillos. Esto se debe a que la red está representada por un grafo dirigido.

En la Figura 4.5 se muestra el segundo conjunto de datos utilizado. Este está completo y contiene aproximadamente 200 nodos, lo que lo hace complicado para manejar. En esta red, los colores de los nodos representan los diferentes grados de vecindad.

La densidad de individuos de grado de vecindad uno es bastante alta. De modo que el radio del anillo uno debe ser mayor que el del resto de anillos, lo que permite acomodar de mejor forma los nodos de una red como la mostrada. En este caso, las distancias entre los nodos están dadas por el numero de mensajes intercambiados, la frecuencia y el grado.

Las pruebas realizadas en esta sección sirven para ilustrar el comportamiento general tanto de los algoritmos clásicos de visualización de grafos como del algoritmo basado en anillos, por esto se utilizó un conjunto de datos reducido con menos de 250 nodos. En la Sección 4.3 se realizarán experimentos con conjuntos de datos de mayor tamaño de tal forma que se pueda evaluar su comportamiento con una mayor carga de datos.

4.3. Pruebas con Redes de Contactos

En la sección anterior se llevaron a cabo algunas pruebas para algoritmos clásicos de visualización de grafos y el algoritmo de fuerza basado en anillos. Tales pruebas buscaban mostrar el comportamiento general de los algoritmos cuando se dispone de un conjunto de datos reducido. En esta sección se desarrollarán pruebas que buscan medir el funcionamiento de los algoritmos desarrollados utilizando

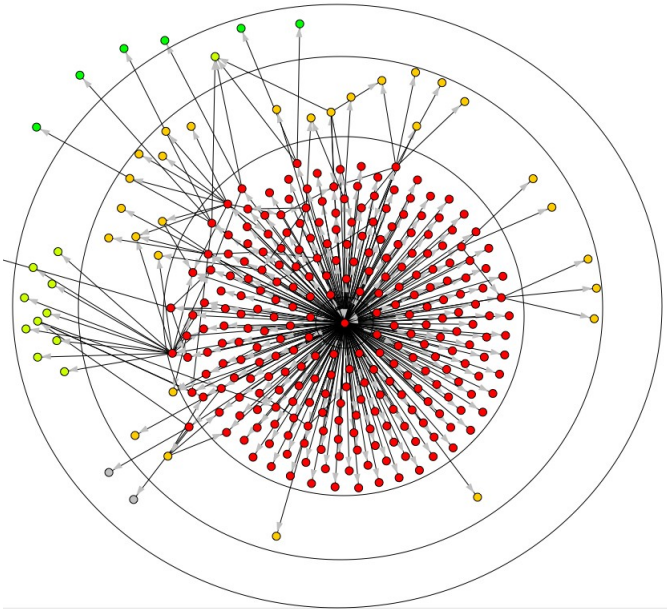


Figura 4.5: Ejemplo del algoritmo de anillos con radio variable utilizando un conjunto de 200 nodos

diferentes conjuntos de datos, los cuales, tienen tamaños diferentes. La idea de las pruebas es mostrar la facilidad que tiene un algoritmo particular para mostrar la información relacionada de los mensajes de correo electrónico. Estas pruebas son ejecutadas utilizando dos conjuntos de datos que corresponden a correos corporativos, el primero, corresponde a la carpeta de entrada de uno de los empleados, por lo que, aunque es más grande que los utilizados en la sección anterior, es el más pequeño utilizado en estas nuevas pruebas, y el segundo, es un conjunto de correos electrónicos que incluye tanto la carpeta de entrada como la de correos enviados.

Primero, se hace una descripción de la configuración experimental, en donde se describen los conjuntos de datos y la forma en que son medidos los resultados. Luego se hacen las pruebas para cada uno de los algoritmos con cada uno de los conjuntos de datos.

4.3.1. Configuración de los Conjuntos de Prueba para los Prototipos

Con el fin de realizar las pruebas relacionadas en este trabajo, se utilizaron dos conjuntos de mensajes, los dos son subconjuntos del conjunto de datos de ENRON, el cual, pertenecía a dicha empresa y fue

liberado una vez finalizó la investigación realizada por la comisión de energía de los Estados Unidos.

En el conjunto de mensajes de ENRON se encuentran 158 usuarios y cerca de 620000 mensajes. Sin embargo, luego de una análisis y una limpieza Klimt & Yang [54] redujeron el número a 200399. Los mensaje eliminados corresponden a información repetida y que a su juicio no aporta información relevante. Para este caso, dado el tipo de información evaluada, si se tienen dos correos con la misma red de contactos y el mismo texto, dichos correos serán tratados como correos iguales y no se perderá información si uno de ellos es eliminado, razón por la cual, la limpieza realizada no afecta los resultados que del sistema presentado se puedan obtener. En general, cada usuario tiene en promedio 757 mensajes.

Para este trabajo se utilizaron dos subconjuntos de correos, cada uno de ellos, perteneciente a un usuario diferente. Cada uno de dichos subconjuntos tiene tamaños diferentes, de modo que se puede

estudiar el comportamiento de cada uno de los prototipos desarrollados en dos escenarios de carga diferentes.

La descripción de cada uno de los conjuntos se muestra en la Tabla 4.1.

Conjunto	Carpetas	Archivos/# de Correos	Nodos
Conjunto I	1	160	448
Conjunto II	116	6071	3174

Tabla 4.1: Descripción de los conjuntos de datos utilizados en las pruebas

La idea de utilizar diferentes tamaños en los conjuntos de prueba está orientada a probar los diferentes algoritmos desarrollados, tal como el de coloreado de nodos y la asociación de nodos con áreas temáticas.

Las pruebas están orientadas a comprobar el funcionamiento de cada uno de los prototipos, comenzando por la generación y pintado de las redes de contactos, luego la generación de áreas temáticas y finalmente la combinación de los dos esquemas de pintado de información.

En primer lugar, se realizarán pruebas con los algoritmos de pintado convencionales, primero Eades y segundo Fruchterman & Reingold. Luego, se presentan algunas pruebas con el algoritmo basado en anillos. Cada prueba contiene además la prueba de pintado de nodos según los grados de vecindad, y, cada una de las pruebas se realiza sobre cada uno de los conjuntos de datos descritos en la configuración experimental.

La medición del desempeño para las pruebas de las redes de contactos será llevada a cabo utilizando algunos de los métodos descritos en [55] y los cuales buscan medir el desempeño de un algoritmo de pintado de grafos a partir de la utilizada y conveniencia para los usuarios de una aplicación particular. A continuación se describen las medidas que se utilizarán para medir la calidad de los grafos:

- Semántica del grafo: esta medida permite calificar un algoritmo de pintado en términos de la utilidad que representa la organización final para una tarea particular para el usuario. En este caso, la identificación de relaciones entre personas y la facilidad para asociar la red de contactos encontrada y los temas de los que pueda conversar en los correos.

- Desempeño de los algoritmos: en este caso se mide el tiempo requerido para efectuar el pintado del grafo. Para esto se supone que las diferentes pruebas son realizadas en la misma máquina y bajo las mismas condiciones.

Es importante anotar que aunque las medidas descritas arriba sirven para comparar el comportamiento de los algoritmos, lo que se busca al final es obtener un método que me permita relacionar el conjunto de personas incluidas en los mensajes de correo y los temas incluidos en dichos correos y mostrar tal relación en el mismo plano.

4.3.2. Experimentos con algoritmos de pintado convencionales

Existen diversos algoritmos para representar grafos, algunos de los cuales han sido explicados en la Sección 4.2. En esta sección de experimentos se mostrarán los resultados de la utilización de los algoritmos de Eades [52] y de Fruchterman & Reingold [2] utilizando la configuración experimental presentada en la Sección .4.3.1.

Primero se presenta el algoritmo de Eades, luego el de Fructerman & Reingold y finalmente el algoritmo de fuerza basado en anillos.

Algoritmo de Eades

Este algoritmo está diseñado para reducir el número de intersecciones que se puedan presentar entre los arcos de la red de contactos. Para esto utiliza la idea de modelar los arcos como resortes que se comportan según la ley de Hooke.

Pruebas con el conjunto de datos I

En la Figura 4.6 se muestra el resultado del algoritmo de Eades para el primer conjunto. Dado que este conjunto solo tiene 160 mensajes de correo electrónico, el número de nodos en la red de contactos es reducido y la ejecución del algoritmo no utiliza mucho tiempo (Este algoritmo es de $O\left(V^2\right)$) y la distribución de los nodos permite identificar algunos grupos de contactos de forma simple.

Adicionalmente, como resultado de la organización, es posible identificar a través de las distancias entre los grupos los nodos que representan a las personas que mayor interacción con el dueño de los mensajes.

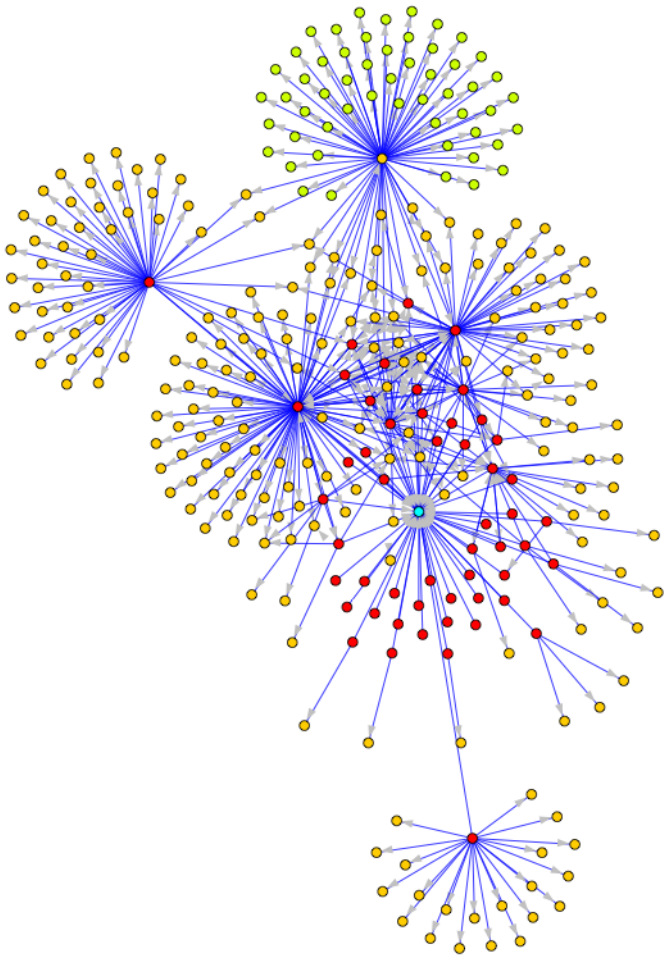


Figura 4.6: Resultado de la ejecución del algoritmo de pintado de EADES para el conjunto de datos I

Con el fin de identificar fácilmente la estructura de la red de contactos, los nodos se colorean según el grado de separación respecto al nodo *ego*.

Según se explicó en la Sección 2.1.1, existen diversos tipos de medidas para las redes sociales. En este caso, la que resulta más interesante es el grado, el cual puede ser calculado de tres formas: (i) grado de entrada, (ii) grado de salida, y (iii) grado consolidado (sumando i y ii). En una red social de este tipo (egocéntrica), es posible determinar el dueño de los mensajes de correos midiendo el valor del grado, ya sea el de entrada, el de salida o el consolidado: el mayor valor señala el dueño.

Sin embargo, en este conjunto particular, para el nodo que representa a la dueña (Tana Jones), se tienen los valores mostrados en la Tabla 4.2.

Grado de Entrada	33
Grado de Salida	0
Grado Consolidado	33

Tabla 4.2: Ejemplo de grados de entrada para las pruebas de pintado del conjunto I

Los cuales no son los valores más altos. Dentro del conjunto existe un nodo con los valores mostrados en la Tabla 4.3.

Grado de Entrada	13
Grado de Salida	84
Grado Consolidado	97

Tabla 4.3: Ejemplo de grados consolidados para las pruebas de pintado del conjunto I

lo que parece contradecir el planteamiento acerca de la relación del valor del grado y los dueños de los mensajes de correo. Pero dado que estos mensajes son un subconjunto de los mensajes de la carpeta de entrada del usuario, se espera que el dueño tenga el mayor grado de entrada, aunque el de salida y el consolidado sean los más bajos, incluso cero.

Cuando se utilizan conjuntos de mensajes completos, es decir, tanto los mensajes recibidos como los enviados, se espera que el nodo que representa al dueño de los mensajes tenga el mayor grado de entrada y de salida, y por consiguiente, el consolidado. Esto se verá en los experimentos con los otros dos conjuntos de mensajes.

Pruebas con el conjunto de datos II

En la Figura 4.7 se muestra la forma en que es pintada la red de contactos del segundo conjunto de datos utilizando el algoritmo de Eades. En este caso se puede ver una gran aglomeración de nodos hacia el centro de la red y algunos grupos en la periferia.

Sin embargo, no es posible identificar información útil de una forma sencilla, por ejemplo, el contacto dueño de los mensajes de correo y sus contactos directos.

Con el fin de identificar el nodo que representa a la persona dueña de los correos es posible determinar el grado de cada uno de los nodos, sin embargo, el algoritmo para calcular el grado de cada uno de los nodos tiene una complejidad muy alta, la cual depende del número de nodos y de vértices existentes. Adicionalmente, si se modifica el tamaño de los nodos en un grafo tan denso, es posible que no se note la diferencia entre dos nodos, por tal razón se aplica el procedimiento de coloreado de los nodos mostrado en la Sección 3.2 el cual es gobernado por la Ecuación 4.5.

En la Figura 4.8 se muestra el resultado del coloreado para el grafo mostrado.

En este caso es posible identificar fácilmente el nodo que representa el dueño de los correos así como los nodos que están directamente conectados. La distribución de los nodos por colores hace que la identificación de conexiones sea más sencilla, aún en redes altamente densas.

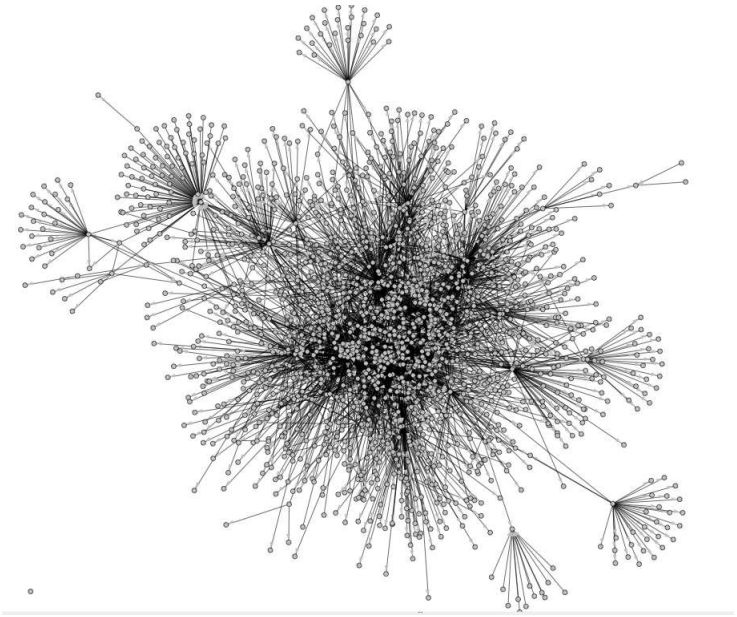


Figura 4.7: Resultado de la ejecución del algoritmo de pintado de Eades para el conjunto de datos II

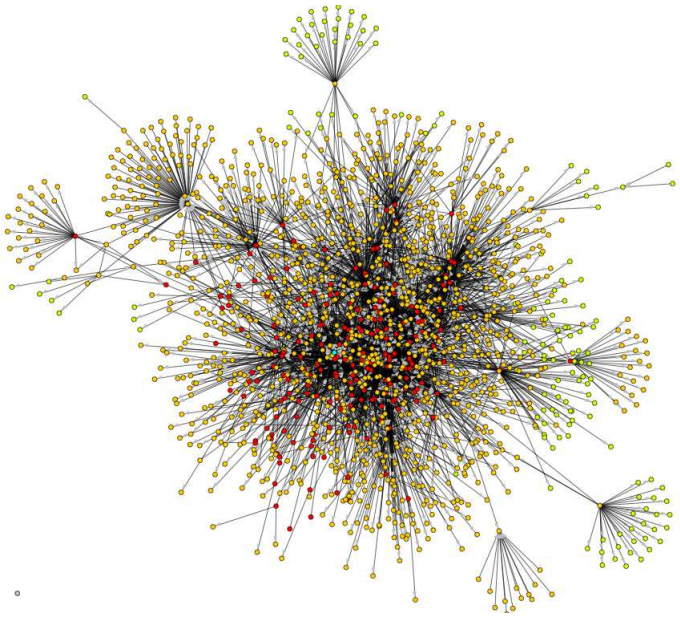


Figura 4.8: Resultado de la ejecución del algoritmo de pintado de Eades para el conjunto de datos II

Análisis General de Resultados del Algoritmo

Este algoritmo tiene un buen desempeño general, es decir, se ejecuta en un tiempo no muy alto, aún para grafos de gran tamaño. Sin embargo, cuando se tienen grafos con muchos nodos y arcos, no es posible determinar a simple vista la estructura de dicho grafo debido al alto grado de aglomeración que

presenta. La evaluación para cada una de las dimensiones descritas arriba se presenta a continuación:

- Semántica del grafo: con pocos nodos, este algoritmo permite identificar la estructura de la red de contactos e identificar los grados a los que se encuentran otros nodos en la red. Sin embargo, cuando existen muchos nodos resulta complicado identificar la estructura de relaciones de la red. Con el fin de obtener más información acerca de la red, es necesario ejecutar procedimiento adicionales, como el de coloreado de los nodos o el del cálculo del grado.

- Desempeño del algoritmo: a medida que el grafo crece la aglomeración de nodos también lo hace, razón por la cual, resulta difícil interpretarla. Adicionalmente, este método de pintado aunque permite los nombres de los correos electrónicos que pertenecen al nodo, no permite determinar los temas del conjunto de mensajes y las relaciones existentes. entre ellos y o nodos. El tiempo de ejecución de este algoritmo es bajo. Sin embargo, cuando el grafo es muy grande, puede ser necesario utilizar procedimientos adicionales, los cueles pueden tomar mucho tiempo en ser ejecutados, por ejemplo, el cálculo del grado de los nodos. En la Tabla 5.9 se muestra un resumen comparativo de los tiempos de ejecución de los diferentes métodos probados.

Algoritmo de Fruchterman & Reingold

Este algoritmo está diseñado para encontrar la mejor distribución posible de los nodos y los arcos dentro de un marco determinado aprovechando al máximo el espacio de pintado.

Pruebas con el conjunto de datos I

En la Figura 4.9 se muestra el resultado de aplicar el algoritmo de Fruchterman & Reingold, en este caso no se observan grupos definidos como los que se definen utilizando el algoritmo de Eades, debido a que este algoritmo no está diseñado para conservar la estructura real de la red, sino optimizar el uso del espacio de pintado.

En este caso también se ha modificado el tamaño de los nodos según el grado que tienen. Para este conjunto de datos, se cambia según el grado de entrada, debido a que el conjunto es tomado de la carpeta de entrada únicamente.

Pruebas con el conjunto de datos II

Este conjunto de datos es el más grande de los dos utilizados para realizar las pruebas, contiene 3174 nodos, lo que hace que se pongan a prueba los algoritmos.

En esta caso el algoritmo intenta distribuir de la mejor manera posible dentro del marco de dibujo todos los nodos de la red. Sin embargo, debido al alto número de nodos, estos tienden hacia el centro de la red social, y de nuevo, solo algunos quedan en la periferia. De igual forma que con el algoritmo de Eades, los grupos más notorios se encuentran en la periferia de la red, sin embargo, dado que no es posible determinar a simple vista que tipos de nodos (conectados directamente al dueño o no) se utiliza de nuevo el procedimiento de coloreado de los nodos, con lo que se obtiene el resultado mostrado en la Figura 4.10.

Análisis General de Resultados del Algoritmo

Este algoritmo tiene un buen desempeño general, aunque su tiempo de ejecución crece considerablemente cuando se organizan grafos de gran tamaño. Adicionalmente, cuando se tienen grafos con muchos nodos y arcos, no es posible determinar a simple vista la estructura de dicho grafo debido al alto grado de aglomeración que presenta. La evaluación para cada una de las dimensiones descritas arriba se presenta a continuación:

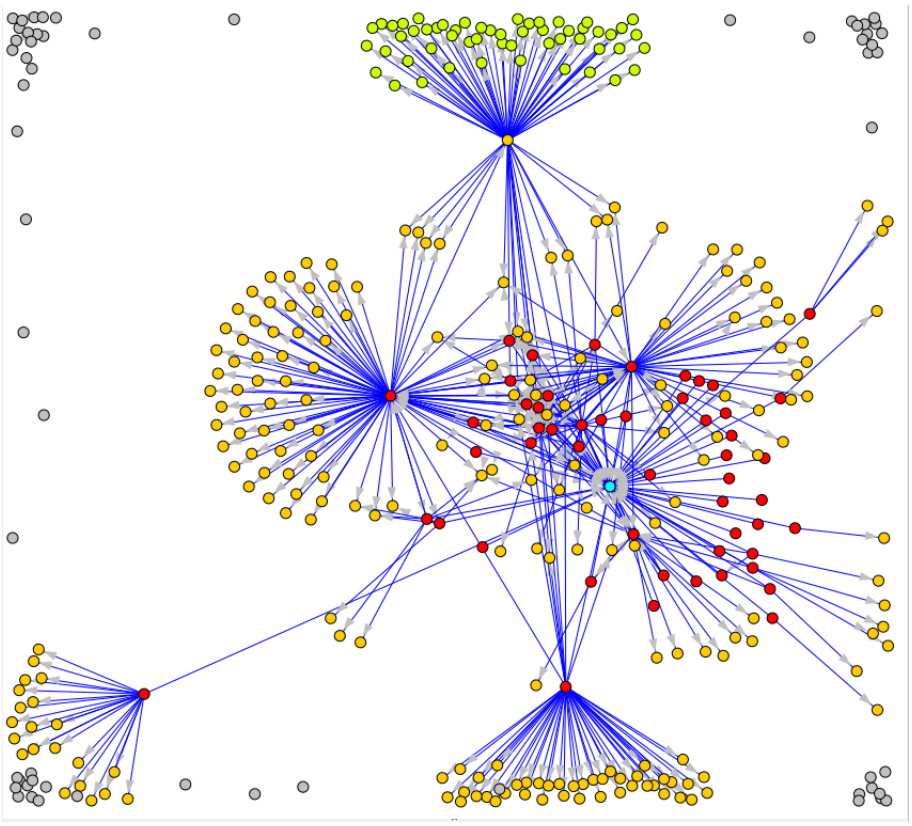


Figura 4.9: Resultado de la ejecución del algoritmo de pintado de Fruchterman & Reingold para el conjunto de datos I

- Semántica del grafo: debido a que este algoritmo está pensado para reducir el número de cruces y de utilizar al máximo el espacio disponible los resultados en términos de visualización son buenos. Esto ocurre aún grafos muy grandes.
  Adicionalmente, permite identificar fácilmente la estructura de la red de contactos e identificar grupos y los diferentes grados a los que puede pertenecer cada uno de ellos.

- Desempeño del algoritmo: este algoritmo al aprovechar el espacio en el va a ser pintado el grafo, muestra grupos de nodos que tienen un grado alto. Esto se debe a que son aquellos nodos los que se encuentran más cercanos al límite del área de pintado. Adicionalmente, este método de pintado aunque permite los nombres de los correos electrónicos que pertenecen al nodo, no permite determinar los temas del conjunto de mensajes y las relaciones existentes entre ellos y los nodos. El tiempo de ejecución de este algoritmo crece notoriamente cuando el tamaño del grafo lo hace lo que lo convierte en un algoritmo un poco pesado. En la Tabla 5.9 se muestra un resumen comparativo de los tiempos de ejecución de los diferentes métodos probados.

4.3.3. Prueba del Algoritmo de Fuerza Basado en Anillos

El algoritmo de fuerza basado en anillos (Sección 4.2.2.2) fue desarrollado para tener una mejor visualización de los diferentes nodos y los grados que ellos tenían. Es decir, que los nodos fueran contenidos por anillos concéntricos según su grado de separación respecto al nodo *ego*. Para esto

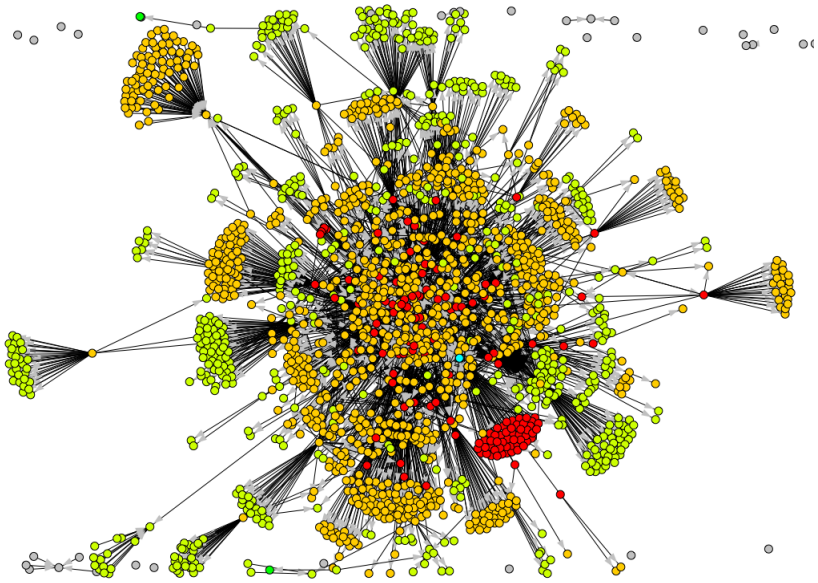


Figura 4.10: Resultado de la ejecución del algoritmo de pintado de Fruchterman & Reingold para el conjunto de datos II con los nodos coloreados.

se asume que los nodos de diferentes grados son fácilmente separables como sucede con la prueba mostrada en las Figuras 4.4 y 4.5.

En este caso, el algoritmo fue probado con los dos conjuntos de datos. Sin embargo, los resultados obtenidos no son satisfactorios debido a la alta interconectividad de los grafos producidos por los conjuntos. Esto quiere decir que algunos de los nodos del grafo están interconectados de tal forma que no es posible encontrar una circunferencia que limite cada uno de los grados de los nodos.

En la Figura 4.11 se muestra el resultado de la ejecución del algoritmo. Se muestra como el anillo agrupa la mayoría de los nodos de grado 1, aunque debido a la alta interconectividad de los nodos de grado 1 y grado 2, no es posible trazar una circunferencia que divida los nodos rojos de los amarillos.

En general, el algoritmo funciona de la forma esperada, separando los nodos de diferentes grados. Sin embargo, presenta algunos inconvenientes cuando existe una alta conectividad de los nodos de diferentes grados, de tal forma que no es posible encontrar una circunferencia que los divida.

Análisis General de Resultados del Algoritmo

Este algoritmo tiene un buen desempeño cuando se dispone de un grafo con ciertas características. Una de dichas características es la de una baja conectividad entre grados de diferentes grados, es decir, que un nodo amarillo no se conecte con varios nodos rojos. Esta restricción no permite que este algoritmo sea utilizado de forma general y por tal razón no es un buen algoritmo de despliegue de grafos.

El tiempo de ejecución de este algoritmo es bajo. Sin embargo, cuando el grafo es muy grande, puede ser necesario utilizar procedimientos adicionales, los cueles pueden tomar mucho tiempo en ser ejecutados, por ejemplo, el cálculo del grado de los nodos.de las dimensiones descritas arriba, suponiendo que se utilice el algoritmo para cualquier grafo, se presenta a continuación:

- Semántica del grafo: debido a que no es posible separar en todos los casos los nodos de diferentes grados, el análisis del grafo se hace complicado.

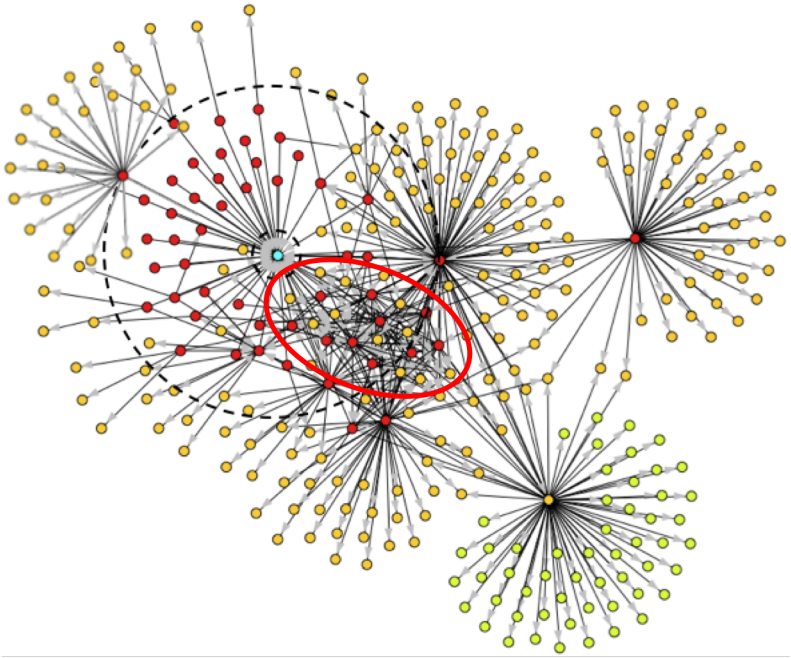


Figura 4.11: Ejemplo de aglomeración de nodos y arcos de grado 1 para el conjunto de datos I

- Desempeño del algoritmo: el tiempo de ejecución de este algoritmo es similar al de Eades, por lo que se ejecuta rápido aún con grafos grandes. En general se tienen tiempos similares a los del algoritmo de Eades, esto es debido a la idea básica utilizada para desplazar los nodos es la misma, solo que existe una restricción dada por los anillos. En la Tabla 5.9 se muestra un resumen comparativo de los tiempos de ejecución de los diferentes métodos probados.

4.3.4. Discusión General de los Algoritmos Presentados

Los algoritmos clásicos de visualización de grafos vistos hasta ahora tiene un buen desempeño general. Adicionalmente, funcionan bien en términos de la intención con la que fueron desarrollados que es mostrar la estructura de la red social. En el caso del algoritmo de Eades y el de Fruchterman & Reingold la idea básica es la de mostrar la red social, su estructura y los grupos que se pueden formar según la afinidad y la configuración de conexiones que presentan.

Con el fin de describir de forma simple la estructura de la red de contactos, se desarrolló el método de coloreado de nodos presentado en la Sección 4.2.2.1,permitiendo determinar fácilmente el tipo de un nodo particular y la separación que tiene del nodo *ego*, sin importar que algoritmo de visualización se utilice, ya que es independiente de el.

Con el fin de simplificar la visualización de la red social, y aprovechando el método de coloreado, se desarrolló el algoritmo de fuerza basado en anillos, el cual está diseñado para agrupar los nodos de diferentes grados en circunferencias concéntricas. Este algoritmo funciona bajo ciertas condiciones particulares, las cuales fueron presentadas en la sección anterior. Sin embargo, su funcionamiento es correcto, agrupando los nodos que cumplen con la restricción de conectividad tal como se muestra en la Figura 4.11. Adicionalmente, el tiempo de ejecución del algoritmo de fuerza basado en anillos tiene un tiempo bajo de ejecución, de hecho similar al de Eades, aún para grafos con más de 2000 nodos.

Aunque todos estos métodos funcionan y cumplen la labor para la que fueron diseñados, todavía no pueden relacionar la información de la red social con los temas e los correos electrónicos, lo cual es

el objetivo final de este trabajo, razón por la cual, se desarrolló un método que permite hacer dicha relación temas – red social, el cual será presentado en el Capítulo 5.

Capítulo 5

Categorización de los Mensajes, Identificación de Temas y Unificación de Esquemas de Representación

"What we have to learn to do, we learn by doing."– Aristotle

La categorización de los mensajes se refiere a la forma en que se agrupan los correos electrónicos según su afinidad temática, es decir, según los diferentes temas que pueden ser tratados por una persona.

Para poder establecer categorías y asociar mensajes de correos a cada una de ellas, es necesario identificar el número de categorías, o conjuntos de temas, existentes en el conjunto de mensajes de correo. Esta identificación se hace a través del análisis de los textos y de la afinidad entre ellos.

Para establecer el número de categorías se utilizan dos técnicas de aprendizaje de máquina: primero, para agrupar los mensajes según su similitud, mapas auto-organizativos, y segundo, para encontrar el número de categorías y los mensajes que a cada una de ellas pertenecen, agrupamiento jerárquico.

5.1. Identificación de Temas

Cada uno de los mensajes del conjunto de mensajes tiene un tema o un asunto, que representa la idea o la intención con la que ha sido enviado. Por ejemplo, un mensaje relacionado con la venta de determinados productos, o la entrega de mercancías, puede estar relacionado con la gestión de los clientes.

De esta forma, se puede decir que un conjunto de mensajes de correo electrónicos tiene asociado un conjunto de categorías de mensajes, las cuales pueden ser utilizadas para agrupar dichos mensajes. Al agrupar los mensajes de esta forma, los mensajes que están más relacionados quedan más cerca y los menos relacionados, estarán más separados.

En la Figura 5.1 se muestra un esquema de la forma en que las categorías podrían ser agrupadas. Cada una de las áreas en la gráfica tiene un color que la identifica, y que es similar al de áreas contiguas, lo que indica que las categorías contienen temas similares o relacionados, pero no iguales.

Esto quiere decir que la cercanía geométrica de las áreas indican también una cercanía en su significado.

Entonces, es necesario encontrar los diferentes temas contenidos en los mensajes, con los cuales, se pueden definir las diferentes categorías representadas en las áreas de modo que se pueda utilizar la distancia entre dichas áreas para relacionar temas similares.

En general, este proceso de identificación de las áreas puede ser llevado a cabo con procedimientos de agrupamiento convencionales o con una combinación de algunos de ellos, como k-means, hierarchical clustering (HC) y self-organizing maps (SOM) entre otros.

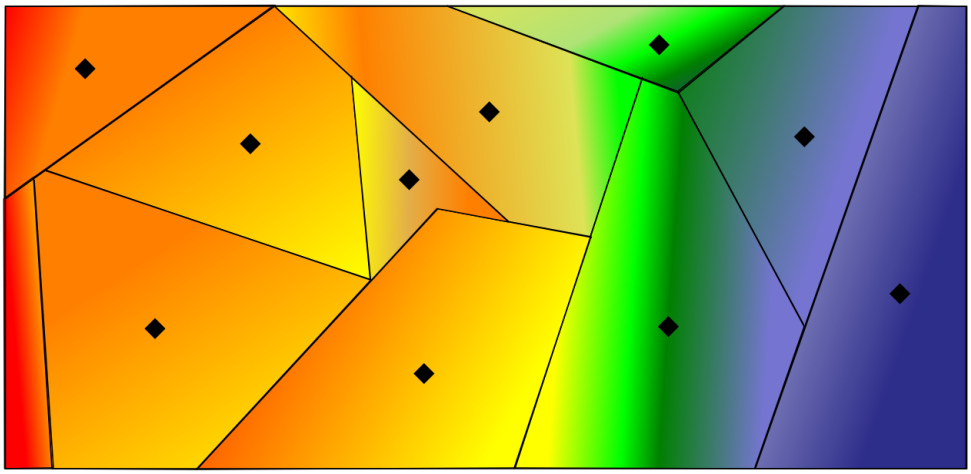


Figura 5.1: Ejemplo de una categorización y las división de las áreas.

K-means: Este método de agrupamiento permite determinar los grupos a los que pertenecen un conjunto de objetos, los cuales en general, son representados como puntos en un espacio $n-$dimensional. Aunque este método tiene buenos resultados en los grupos que produce, el usuario debe ingresar el número aproximado k de grupos que desea obtener y es muy susceptible a los problemas inherentes a la alta dimensionalidad que pueden tener los datos. Para resolver el problema de encontrar un k inicial se han desarrollado algunos métodos pero en general no resulta fácil determinar el número más adecuado.

HC: Este método encuentra los grupos según la distancia (o similitud según sea el caso) entre los objetos a agrupar. Tiene dos variantes, el aglomerativo y el divisivo, el primero, inicialmente trata a cada uno de los objetos como un grupo. Luego, los grupos son mezclados para forma particiones más grandes. Este proceso continúa hasta que se obtiene la partición trivial, es decir, todos los objetos del conjunto pertenecen al mismo grupo. Esta aproximación es desde abajo hacia arriba, se comienza con muchas particiones finas y se obtienen particiones más gruesas [56]. El segundo tipo comienza con un grupo que contiene todos los elementos del conjunto y realiza un procedimiento de partición hasta obtener grupos unitarios.

SOM: Este método permite llevar objetos que se encuentran definidos en espacios $n-$dimensionales a representaciones de 1, 2 o 3 dimensiones, siendo el más utilizado el bidimensional. En el caso bidimensional se utiliza una estructura en forma de matriz, en la cual, cada posición corresponde a una neurona en la red, y cada una de dichas neuronas es probada contra un patrón de entrada. La neurona ganadora será entonces aquella que se encuentre más cerca a dicho patrón, de esta forma se recalculan los pesos de la neurona ganadora y de la vecindad de radio r definida. Este método permite entonces, utilizar un espacio bidimensional para representar grupos que se encuentran en dimensiones más altas, razón por la cual presenta mayor robustez a la alta dimensionalidad de los datos.

Los métodos anteriores, aunque descritos de forma muy general, agrupan otros métodos los cuales son variaciones de estos. En resumen, aunque, el algoritmo de k-means y el de agrupamiento jerárquico presentan buenos resultados, son complicados a la hora de definir los grupos finales requeridos y presentan ciertas debilidades a la hora de trabajar con datos de alta dimensionalidad. Adicionalmente, suponiendo que el problema de la alta dimensionalidad pudiera ser evitado, no es posible de una forma sencilla llevar los grupos encontrados a representaciones visuales sencillas, es decir, llevar los grupos

a espacios de hasta 3 dimensiones, a diferencia de los SOM. Los SOM permiten representar de forma directa grupos de objetos de $n-$dimensiones en espacios de dimensionalidad baja, generalmente 2, lo que permite que la visualización sea sencilla de construir e interpretar. Por esta razón, para representar las categorías de mensajes (los mensajes son objetos representados por vectores $n-$dimensionales), se utilizará un SOM.

La idea de utilizar un SOM en este trabajo es crear un mapa bi-dimensional que represente los temas tratados en el conjunto de correos electrónicos y los muestre como áreas. Estas áreas van a estar formadas por contornos de determinado color y separadas por fronteras de otro color. Con la idea del modelo bi-dimensional es que se plantea la construcción del modelo visual de las áreas que representan los temas de los correos electrónicos. Dicho modelo será explicado más adelante.

5.1.1. Agrupamiento de Mensajes

El primer paso para construir el SOM consiste en tomar el conjunto de correos electrónicos y extraer de cada uno de ellos el cuerpo del mensaje y el asunto. Con esta información, se entrena el SOM, el cual tiene un tamaño de 225 neuronas en una malla de 15×15.

Con el fin de desarrollar este primer paso, se desarrollaran cuatro tareas que se describen brevemente a continuación [57]:

1. Normalización: consiste en eliminar puntuación, eliminar acentos, y convertir todo el texto en minúsculas.

2. Eliminación de *StopWords*: en esta etapa las palabras mas comunes, como artículos, preposiciones, disyunciones y conjunciones son eliminadas del texto.

3. Stemming: cada uno de las palabras del texto es reducida a su forma raíz. Por ejemplo se eliminan la conjugación de los verbos, y se deja la raíz de este. Para este propósito existen muchos algoritmos, entre ellos el de Martin Porter [58]. En este trabajo se utiliza el algoritmo de *snowball* también creado por este autor.

4. Lematización: Las palabras con el mismo significado son cambiadas por una única forma, por ejemplo un sinónimo.

En este trabajo se utilizan la librería "Lucene" [57] para realizar el procesamiento anteriormente descrito. Esta librería ha sido desarrollada dentro de la iniciativa ANT de Apache y está disponible bajo licencia Open Source.

Para la extracción de características, se utiliza frecuencia de términos (TF), la cual es la técnica ampliamente utilizada [25] en categorización de textos. Esta técnica consiste en representar cada documento D_i como un vector que contiene la frecuencia de cada uno de los términos que aparecen en el documento. Formalmente:

$$D_i = \langle F_1, F_2, F_3 ... F_m \rangle \tag{5.1}$$

donde F_1 es la frecuencia del primer término en el documento D_i, y así sucesivamente para cada uno de los términos.

La frecuencia de cada término se calcula de la siguiente manera:

$$F_k = F(t_k, D_i) = \frac{ocurrencias(t_k, D_i)}{N_i} \tag{5.2}$$

donde $ocurrencias(t_k, D_i)$ es el número de veces que aparece el término k en el documento i, y N es el número total de términos en el documento i.

Algunas veces una misma palabra puede presentar una frecuencia elevada no solo en un documento, sino en muchos, por lo tanto no sería una característica representativa de dicho documento. Por lo tanto una manera alternativa de hallar la frecuencia de cada término es:

$$w_{k,i} = F(t_k, D_i) \log \left(\frac{|D|}{|\{D_i \in D : t_k \in D_i\}|} \right) \tag{5.3}$$

donde $|D|$ es el número total de documentos y $|\{D_i \in D : t_k \in D_i\}|$ el número de documentos que contienen al término t_k. En otras palabras, un término es relevante si tiene una alta ocurrencia dentro de un documento y baja en los otros documentos.Esta técnica es denominada *Term Frequency Inverse Document* (TFIDF) [25].

Una de las formas más utilizadas [25] para calcular la similitud de dos documentos representados por vectores es la distancia coseno. Esta medida mide el coseno del ángulo entre dichos vectores:

$$d_{\cos}(D_i, D_j) = \frac{\langle D_i, D_j \rangle}{|D_i||D_j|} \tag{5.4}$$

donde $D_i, D_j \in \mathbf{D}$ son vectores que representan dos documentos.

Luego, con el cuerpo y el asunto del mensaje dispuestos en forma de vector, la red SOM descrita anteriormente es entrenada. Al final, en cada una de las neuronas queda cierto número de correos, los cuales, están relacionados entre si. Esto quiere decir, que el SOM está en capacidad de agrupar los mensajes que tienen cierta relación temática.

Para entrenar la red SOM, cada una de las neuronas del mapa es inicializada con uno de los correos que han sido convertidos en vectores. La selección se realiza de forma aleatoria sobre todo el conjunto de correos.

Una vez se ha inicializado la red, comienza el proceso de entrenamiento, durante el cual, a la red son presentados vectores del conjunto de correos y se calcula la afinidad con cada uno de ellos. La afinidad es calculada con una función de distancia, que en este caso es la distancia coseno que se muestra en la Ecuación 5.4, de tal forma que cada objeto va a estar asignado a la neurona con la que tenga una menor distancia.

Al finalizar el entrenamiento, las neuronas estarán agrupadas según su similitud, los cual es representado por su cercanía geométrica. en el mapa.

En la Figura 5.2 se puede ver un ejemplo del resultado de la ejecución del agrupamiento SOM. Los número en cada uno de las celdas corresponde al número de mensajes que quedaron en la neurona que dicha celda representa.

Adicionalmente, en cada una de las neuronas se muestran 4 etiquetas identificadas para dicha neurona. Para identificar las palabras clave se utiliza el TFIDF, es decir, una palabra será catalogada como clave si su ocurrencia en un documento particular es alta y baja en los otros.

5.1.2. Agrupamiento por Temas

Una vez se dispone del agrupamiento, es necesario identificar las áreas a las que pertenecen los temas incluidos en el conjunto de mensajes de correos electrónicos y adicionalmente, que correos pertenecen a que área. La identificación de las áreas es entonces el ejercicio de encontrar los mensajes en cada una de las neuronas que se relacionan con otros mensajes en otras neuronas.

En la Figura 5.3 se puede ver un ejemplo de un conjunto de neuronas relacionadas. En esta figura se muestran dos áreas arbitrarias de neuronas relacionadas.

Para identificar estas áreas se utiliza el método de clustering jerárquico, de modo que se pueda además, establecer el número final de áreas y asignar a cada una de ellas los mensajes que les corresponden. Cada una de las neuronas tienen una posición (x, y) la cual representa un punto en un espacio bidimensional. Con los puntos de cada área definida, además de definir los bordes, es posible determinar el centro de gravedad, el cual será utilizado posteriormente en el algoritmo de pintado de la red de contactos.

Lo interesante de este método, es que además de definir áreas de mensajes relacionados, la disposición geométrica de las áreas encontradas está relacionada directamente con la similitud de los temas de los correos que se encuentran en cada una de dichas áreas. Esto quiere decir, que entre más cerca

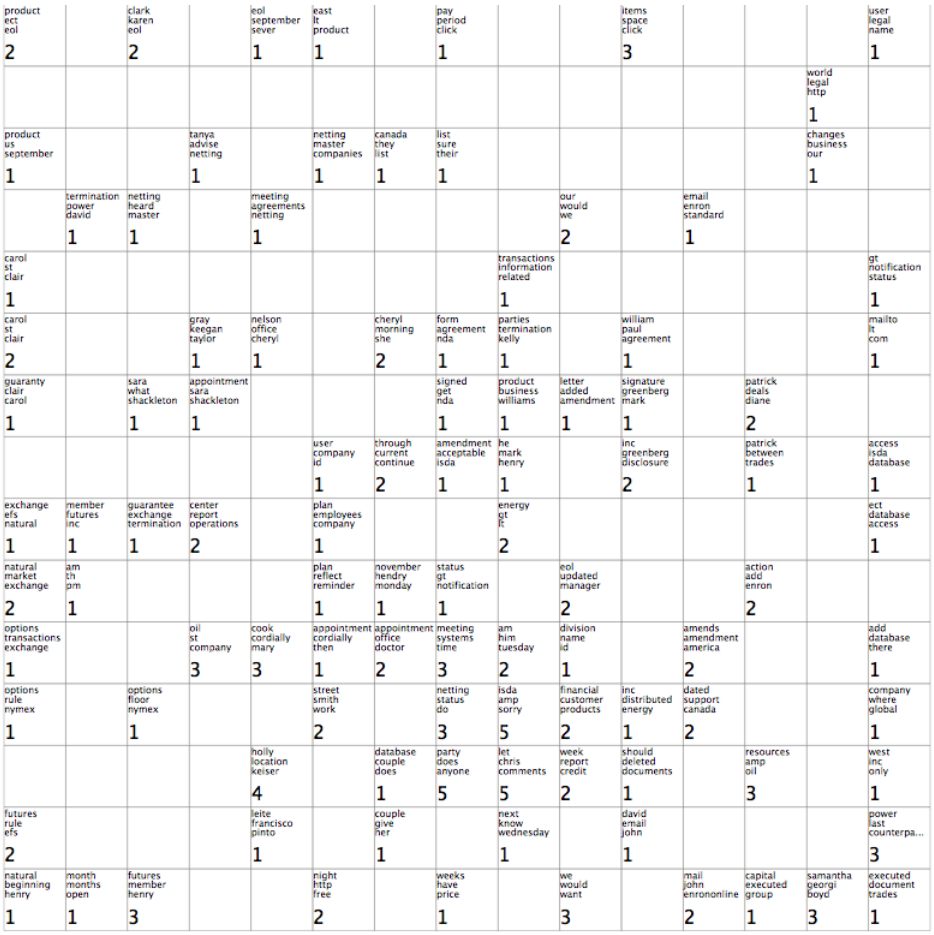


Figura 5.2: Ejemplo del resultado del agrupamiento realizado por el SOM

se encuentren dos áreas, más se van a parecer (no quiere decir que traten los mismos temas) en las temáticas que tratan, o al menos, guardan cierta relación entre ellas.

Una vez se han identificado las áreas, es importante asignarles algún tipo de etiqueta que permita identificar los temas agrupados en ellas. Para esto se utiliza un algoritmo desarrollado por Romero [59] el cual está basado en teoría de la información y en sistemas inmunológicos artificiales.

Este algoritmo identifica las palabras clave de los mensajes que pertenecen a cada una de las áreas sin requerir ningún conocimiento previo del dominio ni tampoco utiliza una *Stop word list*, lo que resulta bastante conveniente a la hora de etiquetar diferentes tipos de temas que pueden llegar a ser muy variados.

Hasta este punto se ha definido el número de áreas sus centroides y los límites respectivos, sin embargo, falta llevarlo a una representación gráfica que sea, de alguna forma, intuitiva para el usuario. En la Sección 5.2 se discute la forma de llevar las áreas encontradas a un esquema visual sencillo.

5.2. Coloreado de las Áreas

En la sección anterior se dijo que era necesaria una representación bidimensional de los grupos de tal forma que se pudiese obtener información geométrica de las áreas y determinar la posición de los

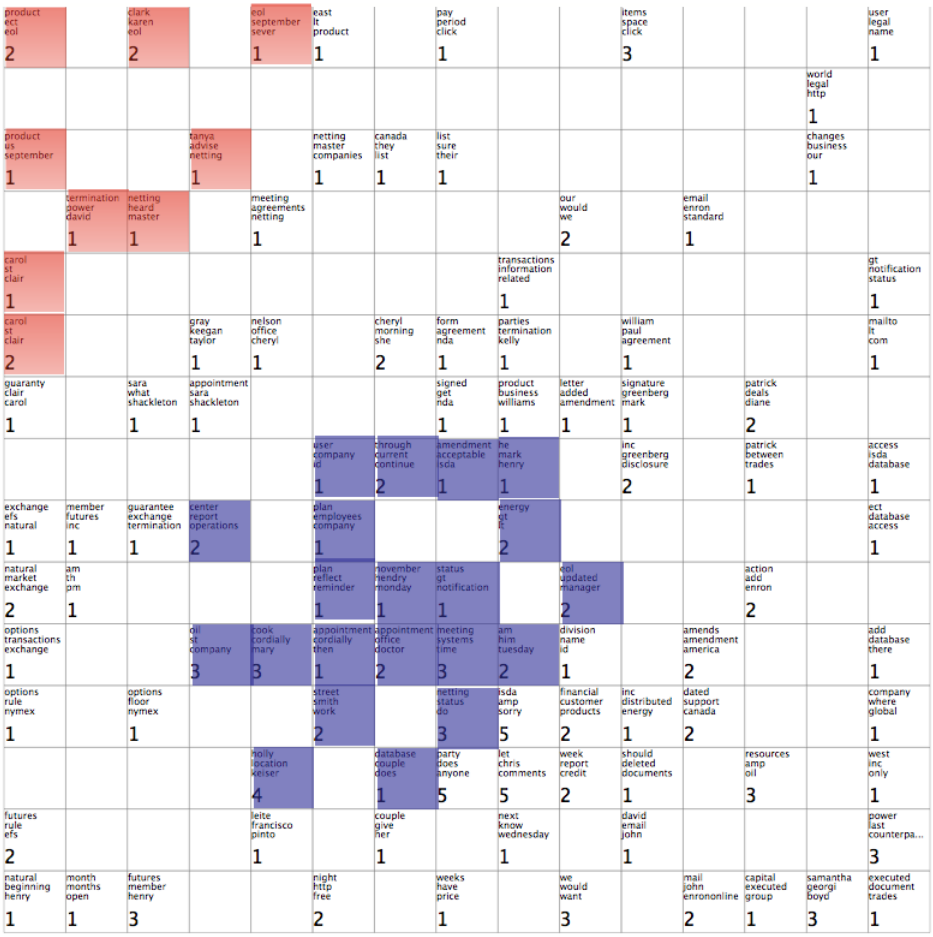


Figura 5.3: Ejemplo de neuronas relacionadas

centroides. Resulta que la representación del SOM ya es bidimensional, por lo que es posible determinar los centroides de las áreas, sin embargo, tiene un inconveniente y es que no permite identificar de forma clara las fronteras existentes entre cada área de los mensajes. Estas fronteras permiten visualizar de una manera más clara tanto el tamaño como la forma de las áreas y así interpretar mejor el alcance de los temas en los correos electrónicos.

De esta forma, una vez se han definido las áreas que definen los temas tratados en los mensajes de correo electrónico, es necesario definir un esquema que permita representar, de una forma sencilla pero que brinde suficiente información, la distribución de los temas encontrados.

Para esto, se define un área que tiene la misma forma de la red SOM utilizada para agrupar los mensajes de correo, la cual en este caso es cuadrada, y para cada uno de los puntos (x, y) de dicha área se determina una intensidad de color según la similitud del punto de prueba, que estará dado por la cercanía al centroide de un área particular. Dicha similitud está dada por:

$$Sim\left(p, C^i\right) = \exp\left(\frac{-\left\|C^i - p\right\|^2}{\sigma^2}\right) \tag{5.5}$$

donde p es el punto de prueba, C^ies el centroide del área $i-$ésima y σ^2 define el radio de cada área y

está determinado por:

$$\sigma^2 = \frac{\#\,mensajes}{k} \tag{5.6}$$

donde k es una constante que define el valor de σ^2.

En la Figura 5.4 se muestra la representación de las áreas de la Figura 5.3 al utilizar la ecuación 5.5.

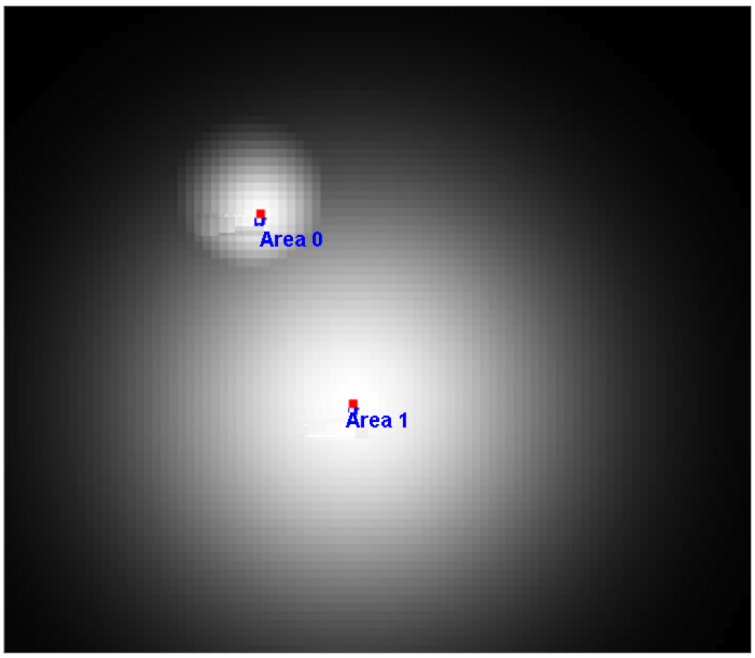


Figura 5.4: Representación de las áreas definidas en la Figura 5.3. Note como las fronteras de las mismas son más claras

Esta representación de las áreas es producto del método de coloreado de áreas desarrollado. En la figura se puede ver la frontera entre las dos áreas, en donde, la intersección entre dichas áreas indica que existe una relación entre los temas de los correos contenidos en cada una de ellas.

5.3. Unificación de los Elementos de Visualización: Pintado de la Red de Contactos Sobre las Áreas

Ya se han definido los elementos necesarios de los que se planea extraer la información. Adicionalmente se han definido los métodos que pueden ser utilizados para representar en un mismo plano la información de los contactos y de los temas categorizados e identificados.

5.3.1. Algoritmo de pintado sobre un campo vectorial de fuerza

En la Sección 4.2 se muestran algunos algoritmos clásicos de pintado de redes sociales (Eades y Fruchterman & Reingold) y una propuesta inicial (algoritmo de fuerza basado en anillos) para facilitar la visualización de las redes sociales. En el primer caso, se tiene que los algoritmos funcionan bien para el propósito con que fueron diseñados, que es el de mostrar la estructura de la red de contactos y las relaciones entre las personas sin tener en cuenta los temas y su distribución. En el segundo caso, el algoritmo funciona bien bajo ciertas condiciones necesarias, que ya fueron presentadas, y que tienen

que ver con la estructura misma del grafo. Sin embargo, el propósito de este trabajo es encontrar la relación entre la red de contactos y los temas existentes en los correos electrónicos.

Por tal razón es necesario desarrollar un método que pueda relacionar los tipos de información. El método desarrollado busca relacionar un conjunto de puntos en un área, los cuales actúan como atractores, con una red social en la que cada nodo tiene un vector e pesos los cuales están relacionados con los puntos atractores.

Cada uno de los nodos de la red, una vez entra al área, se ubica en una posición en la que se encuentre en equilibrio respecto a los puntos atractores.

Esto quiere decir que es posible definir a través del vector de pesos la afinidad de cada nodo con cada uno de los puntos atractores.

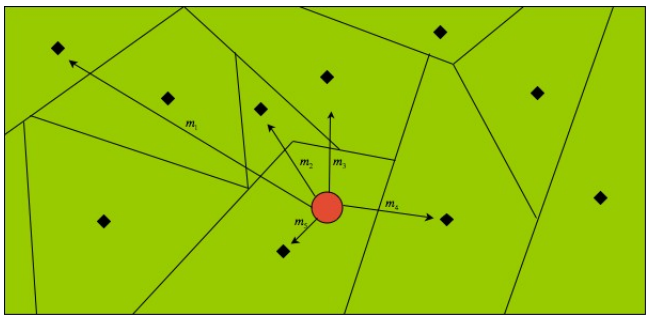


Figura 5.5: Ejemplo de funcionamiento del campo vectorial

La Figura 5.5 muestra un ejemplo de un nodo que entra en el campo vectorial de fuerzas en el que cada uno de los centroides ejerce cierta influencia con intensidad m_i, $i = 1, 2, 3, 4, 5$ sobre dicho nodo.

Entonces, cada nodo tiene un vector $\mathbf{V}$ que define la influencia de cada uno de los centros de gravedad sobre el nodo . Cabe señalar que $\sum_{i \in C} V_i = 1$, donde, C es el conjunto de centroides del área.

Entonces, la posición del nodo está dada por:

$$\begin{aligned} x &= \sum_{i=1}^{|C|} c_i^x V_i \\ y &= \sum_{i=1}^{|C|} c_i^y V_i \end{aligned} \tag{5.7}$$

donde c_i es el centroide $i-$ésimo. Si un centroide determinado no ejerce ninguna influencia sobre un nodo particular, la fuerza será 0: $V_i = 0$.

Con el fin de aplicar este algoritmo es necesario disponer de la información suficiente para construir un área bi-dimensional que contenga un conjunto de centroides (o que al menos sea posible identificarlos) y la información que permita construir un vector de pesos (influencia) para cada uno de los nodos de la red de contactos.

Este modelo de construcción de la visualización permite entonces relacionar en un mismo plano una red de contactos y los temas de los mensajes de correos agrupados de alguna manera. Este aspecto se discutirá más adelante cuando se explique la generación de las áreas y los centroides y la posterior unión de los esquemas de visualización.

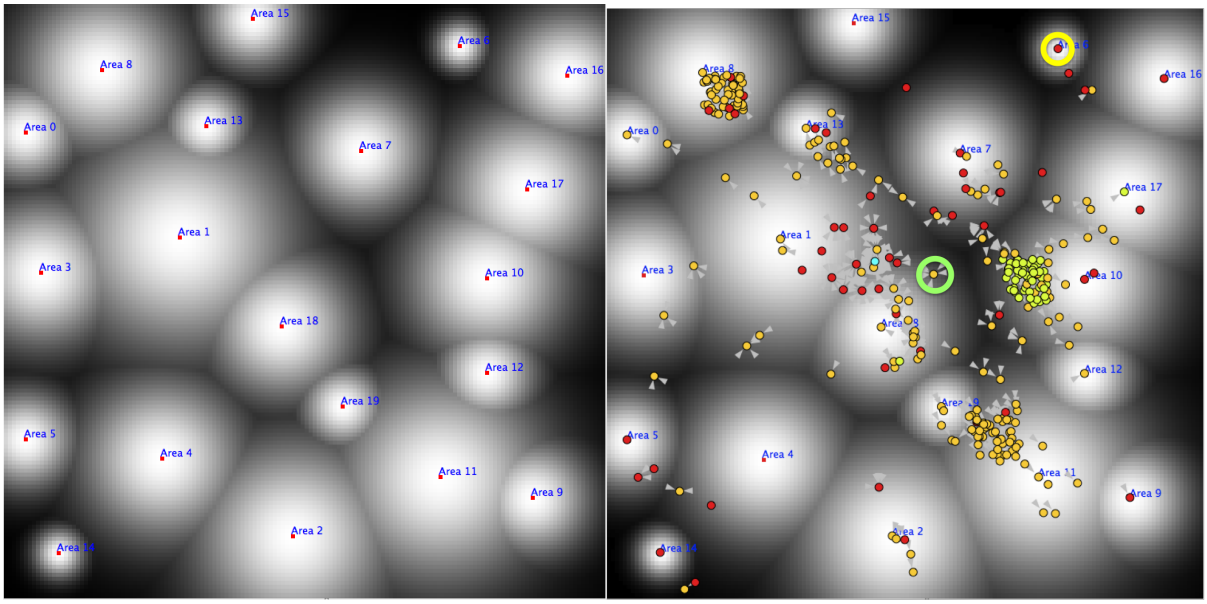
Utilizando la ecuación 5.7 es posible determinar la posición de un nodo dado un conjunto de temas y la afinidad de dicho nodo por cada tema. Sin embargo, es probable que más de un nodo tenga la misma posición, lo que va a llevar a que la red mostrada en el mapa de los temas no refleje la cantidad de nodos reales del conjunto de correos.

Para evitar esto, entonces se la posición solo se asigna la primera vez, y el resto de nodos que pudiesen tener la misma posición, sufrirán una pequeña perturbación, de modo que la nueva posición asignada no esté muy lejos de la calculada con la ecuación 5.7. Entonces, la nueva ecuación de posición queda:

$$Pos = \begin{cases} x = \sum_{i=1}^{|C|} c_i^x V_i, y = \sum_{i=1}^{|C|} c_i^y V_i & si \quad (x,y)\, no\, han\, sido\, asignados \\ x = (1+u)\sum_{i=1}^{|C|} c_i^x V_i, y = (1+u)\sum_{i=1}^{|C|} c_i^y V_i & si \quad e.o.c. \end{cases} \tag{5.8}$$

donde u es una variable aleatoria que se distribuye normal $N \sim \left(C, \frac{a}{3}\right)$, C es la posición del centroide y a es el área geométrica.

En la Figura 5.6 se muestran tanto las áreas identificadas como la forma en que los nodos son localizados en las mismas áreas utilizando la Ecuación 5.8. Note en la Figura 5.6(b) existen algunos nodos que están localizados justo encima del centro de masa de un área. Este es el caso de del nodo encerrado en el círculo amarillo, el cual, se encuentra sobre el centro de masa del área 6. Estas áreas son obtenidas a partir del conjunto de datos I (ver Sección 4.3.1).



(a) Ejemplo de áreas identificadas para un conjunto particular de correos electrónicos

(b) Ejemplo de áreas identificadas para un conjunto particular de correos electrónicos y la red social asociada a él

Figura 5.6: Ejemplo de las áreas temáticas identificadas y de la red social asociada. Cada uno de los nodos está localizado según el valor de afinidad con los temas.

El nodo seleccionado pertenece a un usuario cuyo nombre es Kay Young y se comunica directamente con el dueño de los correos (Tana Jones) para enviarle un correo respuesta cuyo asunto es: *RE: Russian NDAs*.

Nótese que la varios nodos rojos (los que tienen conexión directa con el dueño de los correos) están ubicados cerca de los centroides de diferentes áreas. Estos nodos corresponden a persona con las que el dueño de los correos intercambia comunicaciones sobre temas particulares de forma directa. Sin embargo, existen nodos rojos que se encuentran ubicados sobre las áreas negras, las cuales son las

fronteras en tres los temas. Esto quiere decir que las personas representadas por estos nodos, aunque tienen conexión directa con el dueño, participan de diversas comunicaciones y, posiblemente, no tengan mucho peso en un solo tema determinado.

5.4. Pruebas con la Identificación de Temas

Los experimentos relacionados con la identificación de las áreas que representan los temas de los mensajes de correo están pensadas para probar el funcionamiento del segundo prototipo. Para estas pruebas se utiliza una configuración experimental similar a la utilizada en la Sección 4.3.1, pero se incluye un conjunto de datos nuevo, el cual está compuesto de 16 carpetas, 3265 mensajes de correo y 1561 nodos. Entonces, la nueva configuración experimental queda como se muestra en la tabla 5.1.

Conjunto	Carpetas	Archivos/# de Correos	Nodos
Conjunto I	1	160	448
Conjunto II	116	6071	3174
Conjunto III	16	3265	1561

Tabla 5.1: Descripción de los conjuntos de datos utilizados en las pruebas de identificación de temas y áreas.

5.4.1. Pruebas para seleccionar el número de áreas

Las pruebas relacionadas con el número de áreas fueron ejecutadas utilizando los cuatro conjuntos descritos en la Sección 4.3.1. Tales pruebas fueron realizadas de la siguiente forma: para cada conjunto de datos se definieron cuatro pruebas para tres valores fijos de áreas: 10, 15 y 20, y finalmente se dejó que el algoritmo de agrupamiento jerárquico encontrara el valor óptimo de áreas a partir del coeficiente de silueta. El coeficiente de silueta combina los conceptos de cohesión y separación de los grupos para definir la estructura de los mismos [60]. Es decir, busca que los puntos de un grupo estén muy cerca, pero la distancia entre grupos sea alta.

Para calcular el coeficiente de silueta primero se calcula la distancia promedio a_i del objeto $i-$ésimo de un grupo particular respecto a los otros objetos en el grupo, luego, se calcula la distancia promedio b_i del mismo objeto $i-$ésimo respecto a los elementos en otros grupos. El coeficiente s_i estará dado por:

$$s_i = \frac{b_i - a_i}{\max(a_i, b_i)}$$

El valor del coeficiente puede variar en el intervalo $[-1, 1]$. Entonces, una medida de la calidad del agrupamiento es tomar el valor promedio de los coeficientes de silueta de cada grupo y luego encontrar el valor promedio de todos los grupos, el cual, debe ser cercano a 1.

Dado que se utiliza un algoritmo de agrupamiento jerárquico aglomerativo, es posible tener tantos grupos como correos existan, es decir, tratar cada uno de los mensajes como un tema particular.

La idea general es la de reducir el número de intersecciones de las áreas encontradas. De esta forma, se pueden obtener únicamente las áreas que contengan información diferente entre si, es decir, que la diferencia entre los temas de dos áreas determinadas sea significativa.

Debido a que se utilizan un configuración igual para cada conjunto de datos, en primer lugar se van a mostrar los resultados de la variación de áreas para el conjunto de datos II que es que mayor correos y nodos tiene. Luego se presentan los resultados con el cálculo de áreas con el número óptimo para cada uno de los tres conjuntos de datos.

Las pruebas de identificación de las áreas consisten en identificar los temas tratados en los correos. La primera prueba consiste en generar un número alto de áreas para cada uno de los conjuntos de

datos. En la Figura 5.7 se muestran los resultados para cada uno de los conjuntos de datos cuando se seleccionan 20 temas.

En este caso se evidencian un conjunto de áreas pequeñas en torno a áreas más grandes. Dichas áreas pequeñas están intersecadas con las más grandes lo que quiere decir que existe un conjunto de correos que tratan ambos temas, o que, pueden ser categorizados en cualquiera de ellos.

Es importante notar que la mayoría de áreas mostradas en los ejemplos presentados en las Figuras 5.7(a), 5.7(b) y 5.7(c) son de un tamaño pequeño, lo que quiere decir que no contienen un número significativo de mensajes de correo electrónico en ellas. Dichas áreas pequeñas indican que es posible reducir el número de áreas, es decir, que en todo el conjunto de correos podrían agruparse los temas tratados en menos áreas. De esta forma, en la Figura 5.7(d) se ve como la reducción del número de áreas reduce también el número de intersecciones presentadas.

Finalmente, se realizan las pruebas en las que el algoritmo de agrupamiento jerárquico encuentra el número óptimo de áreas en cada uno de los conjuntos de datos. En la Figura 5.8 se muestran los resultados obtenidos.

En este caso se para todos los conjuntos de datos se obtienen cuatro áreas bien definidas. El hecho que con los tres conjuntos de datos se obtengan cuatro áreas no se puede tratar como una coincidencia; los conjuntos de mensajes, aunque son de diferentes usuarios, son de la misma compañía.

El número óptimo de áreas parece reducido, sin embargo, agrupa de la mejor forma los diferentes temas de tratados en las conversaciones de los correos electrónicos.

El conjunto I está dividido en cuatro áreas, de las cuales el área 0 es la que mayo número de correos tiene. El área 0 agrupa mensajes que tratan temas relacionados con los negocios asociados a la producción de energía y de derivados del petróleo. El área 1 contiene temas relacionados con planes de expansión y exploración petrolera. El área 2 tiene correos relacionados con transacciones en la bolsa de Nueva York y de otros negocios de alto nivel de la compañía. El área 3, trata mensajes varios relacionados con el día a día de la compañía, incluyendo algunos mensajes de índole personal.

En la Tabla 5.2 se muestran algunas de las palabras clave para cada una de las áreas del conjunto I. Estas palabras están organizadas según la frecuencia de aparición en cada una de las áreas.

El conjunto II también está dividido en cuatro áreas, de las cuales, el área 3 es la que tiene el mayor número de mensajes de correo. El área 0 contiene mensajes relacionados con producción y distribución de energía no asociada únicamente con el petróleo. El área 1 contiene mensajes relacionadas con el mercado energético, incluyendo temas de venta y compra de energía. El área 2 contiene correos relacionados regulación y legislación energética. El área 3, contiene mensajes de correo relacionados con mensajes internos de la compañía y del día a día, así como mensajes relacionados con negociaciones de energía en bolsa.

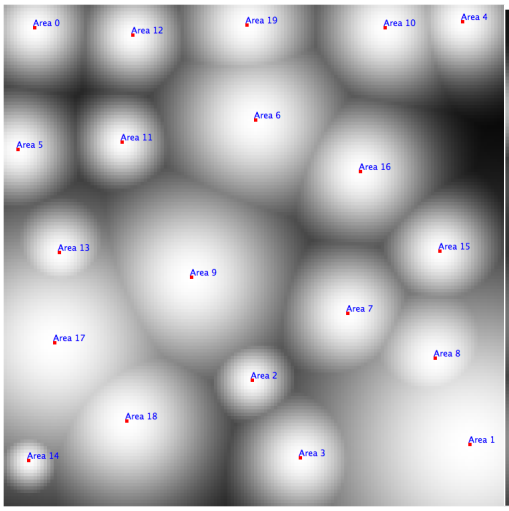
En la Tabla 5.3 se muestran algunas de la palabras clave de cada una de las áreas del conjunto II.

Al igual que los primeros conjuntos, el conjunto III también está dividido en cuatro áreas. En el área 0 están mensajes relacionados con información corporativo interna y mensajes personales. El área 1 contiene mensajes relacionados con el mercado energético. El área 2 contiene mensajes relacionados con el día a día de la compañía. El área 3 contiene correos en su mayoría personales.

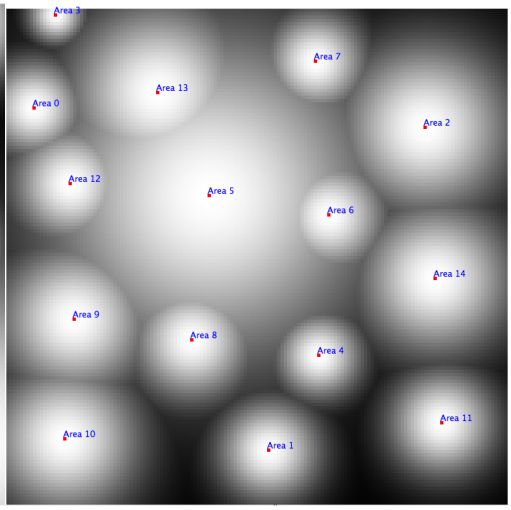
En la Tabla 5.4 se muestran algunas de las palabras clave contenidas en cada una de las áreas temáticas del conjunto III. Estas palabras están organizadas según su frecuencia de aparición en los mensajes en cada una de las áreas identificadas. Adicionalmente para todos los conjuntos, se muestra el número de correos contenidos en cada una de las áreas de cada conjunto. Note que efectivamente las área de mayor tamaño contienen el mayor número de mensajes de correo electrónico.

5.5. Experimentos con el Método Unificado

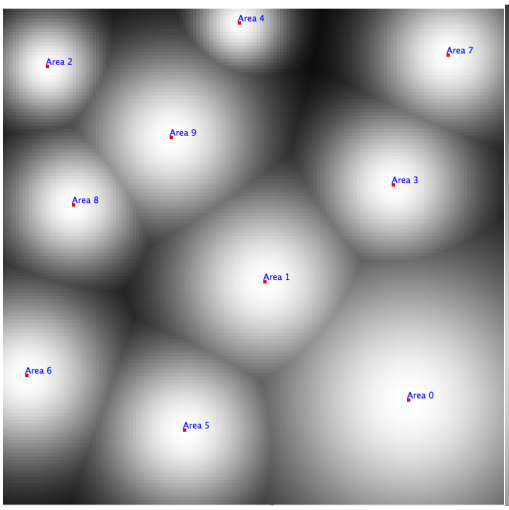
Una vez se han encontrado las áreas en las que se pueden agrupar los temas contenidos en los correos electrónicos, es posible utilizar el método definido en la Sección 5.3.1. En esta sección se presentan los experimentos realizados para medir el desempeño del algoritmo basado en campos de fuerza, el



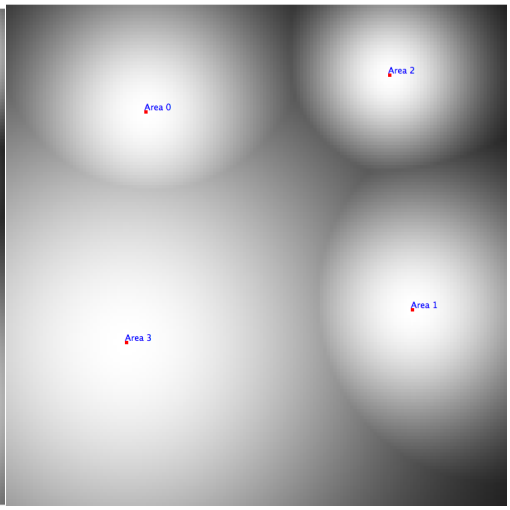
(a) Muestra de las veinte áreas encontradas para el conjunto de datos III



(b) Ejemplo de las quince áreas encontradas para el conjunto de datos III

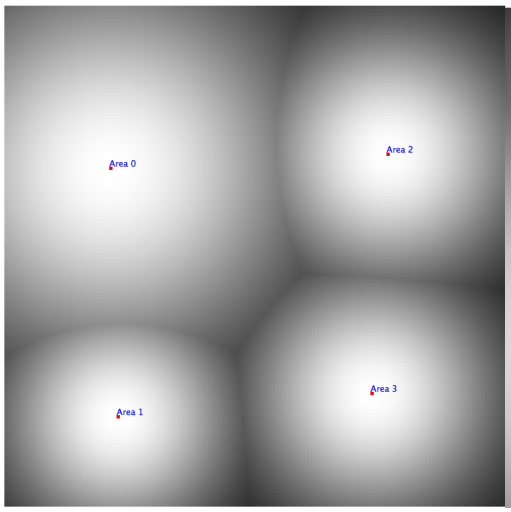


(c) Ejemplo de las diez áreas encontradas para el conjunto de datos III

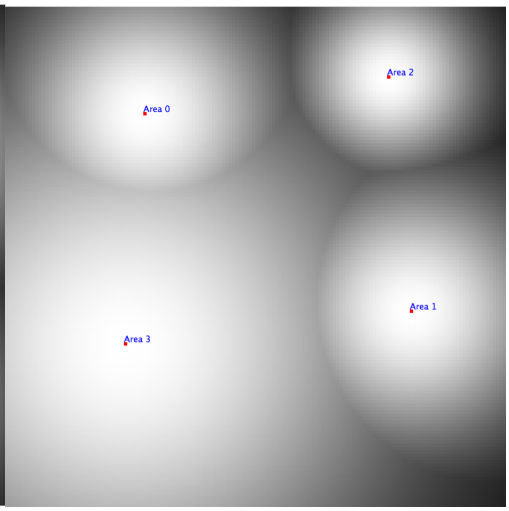


(d) Ejemplo de las cinco áreas encontradas para el conjunto de datos III

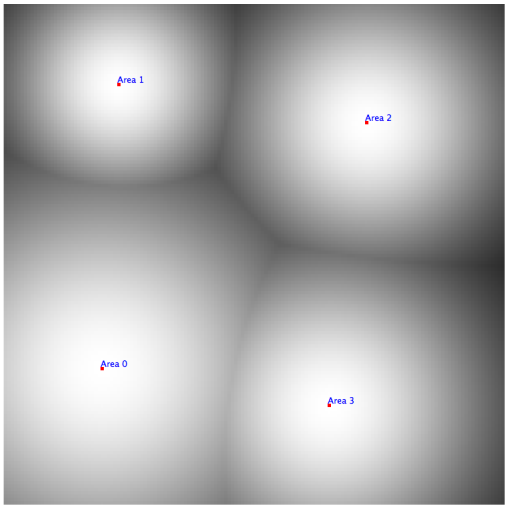
Figura 5.7: Comparación de las áreas encontradas para el conjunto III



(a) Número óptimo de áreas encontardas para el conjunto de datos I



(b) Número óptimo de áreas encontardas para el conjunto de datos II



(c) Número óptimo de áreas encontardas para el conjunto de datos III

Figura 5.8: Áreas encontradas para los conjuntos de datos I, II y III. Este número de áreas es determinado utilizando el coeficiende silueta durante la ejecución del algoritmos de agrupamiento jerárquico.

cual, puede relacionar la información de las redes sociales con la información temática incluida en los mensajes de correo.

Primero se presenta la configuración experimental utilizada, luego se presentan los resultados obtenidos y finalmente se presenta una discusión y análisis de los mismos.

5.5.1. Configuración de los Conjuntos de Prueba para el Prototipo

Con el fin de realizar las pruebas relacionadas en este trabajo, se utilizaron dos conjuntos de mensajes, los dos son subconjuntos del conjunto de datos de ENRON, el cual, pertenecía a dicha empresa y fue liberado una vez finalizó la investigación realizada por la comisión de energía de los Estados Unidos.

En el conjunto de mensajes de ENRON se encuentran 158 usuarios y cerca de 620000 mensajes. Sin embargo, luego de una análisis y una limpieza Klimt & Yang [54] redujeron el número a 200399. Los mensaje eliminados corresponden a información repetida y que a su juicio no aporta información relevante. Para este caso, dado el tipo de información evaluada, si se tienen dos correos con la misma red de contactos y el mismo texto, dichos correos serán tratados como correos iguales y no se perderá información si uno de ellos es eliminado, razón por la cual, la limpieza realizada no afecta los resultados que del sistema presentado se puedan obtener. En general, cada usuario tiene en promedio 757 mensajes.

Para este trabajo se utilizaron tres subconjuntos de correos, cada uno de ellos, perteneciente a un usuario diferente. Cada uno de dichos subconjuntos tiene tamaños diferentes, de modo que se puede estudiar el comportamiento de cada uno de los prototipos desarrollados en dos escenarios de carga diferentes.

La descripción de cada uno de los conjuntos se muestra en la Tabla 5.5.

La idea de utilizar diferentes tamaños en los conjuntos de prueba está orientada a probar los diferentes algoritmos desarrollados, tal como el de coloreado de nodos y la asociación de nodos con áreas temáticas.

Las pruebas están orientadas a comprobar el funcionamiento de cada uno de los prototipos, comenzando por la generación y pintado de las redes de contactos, luego la generación de áreas temáticas y finalmente la combinación de los dos esquemas de pintado de información.

En primer lugar, se realizarán pruebas con los algoritmos de pintado convencionales, primero Eades y luego Fruchterman & Reingold. Luego, se presentan algunas pruebas con el algoritmo basado en anillos. Cada prueba contiene además la prueba de pintado de nodos según los grados de vecindad, y, cada una de las pruebas se realiza sobre cada uno de los conjuntos de datos descritos en la configuración experimental.

La medición del desempeño para las pruebas de las redes de contactos será llevada a cabo utilizando algunos de los métodos descritos en [55] y los cuales buscan medir el desempeño de un algoritmo de pintado de grafos a partir de la utilizada y conveniencia para los usuarios de una aplicación particular. La medición del desempeño del método se hará de la misma forma planteada en la Sección 4.3.1.

5.5.2. Resultados

Luego de encontrar las áreas que definen los temas contenidos en cada uno de los conjuntos de correos electrónicos utilizados para las pruebas, es posible utilizar la Ecuación 5.8 para ubicar la red de contactos contenida en cada uno de dichos conjuntos de mensajes.

Al ubicar los nodos de esta forma es posible relacionar a cada uno de los actores (personas e instituciones) con los temas que se encuentran en los mensajes de correo. La fuerza de la relación entre una persona y un tema entonces está dada por la cercanía del nodo que representa a esa persona y el centroide del área que representa el tema.

En la Figura 5.9 se muestra el resultado de la ejecución del algoritmo de pintado basado en campos de fuerza. Se ven dos grandes grupos de nodos, uno, cerca al centro de una de las áreas (Área 3) y

otro, hacia el centro de la gráfica, en donde además se encuentra el nodo que representa al dueño de los mensajes. Los nodos que se encuentran en el centro son aquellos que hablan de todos los temas.

Otros nodos se encuentran cerca a la frontera de dos áreas. Dichos nodos representan actores que solo hablan de dos temas. Note que debido a que este tipo de redes puede contener un número alto de arcos, o relaciones, la visualización puede resultar compleja para los usuarios. Por esta razón se desarrolló una versión del prototipo en el que las relaciones no se visualizan, sino que es posible elegir un nodo al que sus arcos serán pintados. Los arcos pintados son tanto los entrantes como los salientes.

Esta variación en el modelo de pintado además de permitir que los arcos no cubran las áreas, permite que se puedan identificar las relaciones entre nodos específicos, es decir, que se puedan identificar las relaciones entre nodos al no presentarse la aglomeración de arcos.

En la Figura 5.5.2 se muestra el resultado de la ejecución del algoritmo de unificación de métodos de visualización para el conjunto II. Adicionalmente, se muestra la comparación entre realizar la uni3ficación mostrando todos los arcos, Figura 5.10(a), y, sin mostrar ningún arco de la red, Figura 5.10(b).

El conjunto II es significativamente mayor que el primero, razón por la cual se va una aglomeración mayor de nodos en torno a las áreas. En este caso los grupos más grandes se ven cerca de los centroides de las áreas, lo que indica que existen personas con las que el usuario trata únicamente ciertos temas.

De nuevo, el usuario dueño de los correos está cerca al centro geométrico de las áreas; realmente se encuentra más cerca a una de las áreas, lo que indica que dicha área puede estar relacionada con su trabajo diario.

Finalmente, en la Figura 5.11 se muestra el resultado de la ejecución del algoritmo para el tercer conjunto de datos. En este caso la aglomeración de nodos es más notoria debido al tamaño del grafo.

Note como al seleccionar los arcos del nodo de grado 1 es posible hacer un seguimiento más sencillo a las relaciones existentes. Adicionalmente en la Figura 5.11 se resaltan tres nodos, dos de ellos en dos de las áreas, y uno en la frontera de dos áreas de mensajes; Todos los nodos seleccionados son de grado 1. la información de cada uno de dichos nodos es resumida en las Tablas 5.6 y 5.7 a continuación.

El correo resumido en la Tabla 5.6 pertenece al área 0 del conjunto II, la cual trata temas relacionados con acuerdos de distribución y producción de energía. Este correo es un mensaje relacionado con la finalización de acuerdos con el gobierno, pero es más de gestión de proyectos y contratos.

El correo resumido en la Tabla 5.7 pertenece al área 3 del conjunto II. En esta área se encuentran contenidos correos relacionados con el día a día de la compañía. En este caso, el correo está relacionado con un proyecto interno que será ejecutado.

Para el uso de este algoritmo de pintado es importante asignar colores a los nodos. Como se ve en el resultado de la ejecución para el conjunto II (Figura 5.11) la aglomeración cerca de los centroides puede ser tal que sea imposible obtener información importante acerca de la estructura de la red.

El nodo que se encuentra señalado con la etiqueta *0-3*, que se encuentra entre las áreas 0 y 3 comparte información de cada una de dichas áreas. El resumen del contenido de los mensajes de correo se presenta en la Tabla 5.8.

En el caso del primero correo presentado en la Tabla 5.8, el cual pertenece al área 0 del conjunto II, se tiene un mensaje relacionado con la negociación de energía, pero no con terceros sino en términos de requerimientos de mediación. En el caso de los correos 2 y 3, que pertenecen al área 3, están relacionados con negociaciones de producción y venta energía de todo tipo en la bolsa de Nueva York.

En este caso, es importante resaltar que aunque el método desarrollado está en capacidad de identificar la división de los temas y de agrupar los mensajes de correo según tal distribución, no está en capacidad de identificar el contexto completo del mensaje.

Finalmente, en la Figura 5.8 se presenta el resultado de la unificación de las áreas temáticas y las redes sociales para cada uno de los conjuntos. En este caso los colores de los nodos permiten identificar la estructura de la red en términos de los grados de separación de cada uno de dichos nodos respecto al nodo que representa al dueño de los correos.

Note como los nodos no están concentrados en los centros de masa de las áreas, sino que se encuentran por todo el mapa, mucho de ellos cerca al nodo *ego*. En el centro del mapa se encuentran las personas que manejan varios temas de los que el dueño de los correos trata y pueden ser fuentes

de consulta para dudas básicas. Sin embargo, si se quisiera buscar una persona experta, o que tenga cierto conocimiento avanzado en un tema específico, se deben buscar personas que se encuentren cerca del centro de masa del área que contiene el tema buscado.

A continuación se presenta una discusión acerca de los resultados obtenidos al utilizar el método propuesto para visualizar la información de los correos electrónicos.

5.5.3. Análisis de Resultados

Utilizando la misma metodología planteada en la Sección 4.3 se mide el desempeño general del algoritmo planteado. A continuación se presenta un resumen de la evaluación de los algoritmos de pintado de grafos:

- Semántica del grafo: este algoritmo permite a los usuario identificar la asociación entre personas y temas contenidos en los mensajes de correo electrónico. Adicionalmente le permite ver los correos mismos que una persona dentro de la red ha enviado y/o recibido.
- Desempeño de los algoritmos: aunque este algoritmo permite identificar fácilmente la asociación entre personas y temas, hace mucho nodos se agrupen en torno al centroide de un área particular. Esto pude en un caso determinado aumentar la complejidad para leer la información allí contenida. Debido a la forma en que se calculan las posiciones de los nodos en el mapa, la complejidad del algoritmo es directamente proporcional al número de áreas encontradas y al número de nodos, lo que lo hace eficiente en tiempo.

En general, este método permite mostrar en el mismo plano la agrupación y distribución de los temas tratados en los mensajes de correo electrónico así como el conjunto de contactos incluidos en los mensajes electrónicos del conjunto. Es posible además ver la estructura de la red a través de los arcos que conectan los nodos del grafo, sin embargo, en un conjunto de datos grande como el II, la aglomeración de arcos no permite ver claramente la ubicación de algunos nodos y las fronteras de las áreas. Por esta razón, los arcos de la red se esconden, de tal forma que se puedan elegir algunos para mostrar, y se utilizan los colores definidos en el coloreado de nodos para identificar el área de la red social.

En la Tabla 5.9 se muestran los tiempos de ejecución de los algoritmos para los dos primeros conjuntos utilizados en las pruebas. Note la diferencia en el tiempo de ejecución de los algoritmos propuestos respecto a los algoritmos clásicos. Además de la diferencia de ejecución de tiempo, la principal ventaja del método basado en campos de fuerzas es la capacidad que tiene para relacionar la red de contactos con los temas del correo electrónico.

Es importante anotar que los algoritmos clásicos desempeñan muy bien las tareas para las que fueron diseñados, que son las de presentar la red social y su estructura con fines de análisis de conectividad y otros de presentación, pero no para asociar dicha red con otra información, la cual no puede ser identificada a simple vista.

AREA	PALABRAS CLAVE CONJUNTO I	CORREOS ASOCIA-DOS
0	meeting, requested, canada, amendment, does, week, company, report, america, credit, nda, natural, isda, file, agreements, netting, work, appointment, disclosure, rohauer, samantha, wanted, landau, action, center, patrick, status, gas, smith, operations, signed, exchange, user, market, energy, office, master, greenberg, street, georgi, enron, add, notification, boyd	55
1	product, futures, company, agreement, mark, file, cook, database, mary, plan, users, cordially, efs, access, canada, exchange, landau, rule, netting, samantha, does, georgi, transaction, boyd	33
2	eol, click, space, limit, items, manager, futures, updated, information, access, deals, file, status, list, patrick, does, moran, notification, clark, brackett, diane, karen, nymex, counterparty, months	37
3	keiser, file, party, oil, does, holly, department, location, office, morning, wednesday, cheryl, clair, comments, directly, appointment, meeting, requested, doctor, know, carol, resources, agreement	35

Tabla 5.2: Palabras clave de cada área del conjunto I

ÁREA	PALABRAS CLAVE CONJUNTO II	CORREOS ASOCIA-DOS
0	draft, enron, state, attached, doc, epsa, committee, board, energy, charles, mail, working, article, ferc, customers, electricity, nerc, analysis, comments, thank, filed, enronxgate, org, affairs, prices, texas, year, dasovich, states, power, new, confidential, legislation, intended, sarah, ees, management, utility, rto, rates, legislative, report, industry, language, past, rate, want, program, market, linda, novosel, christi, davis, contracts, price, jeff, brief, california, summary, daily, reports, michael, europe, based, announcement	1306
1	enron, news, status, http, europe, chairman, html, new, report, employee, business, said, meeting, told, june, lay, group, ferc, senate, million, european, dynegy, jones, corp, markets, customers, crisis, credit, management, provide, risk, market, services, nicolay, california, talking, service, changes, ebs, board, deal, action, hearing, make, development, bush, christi, costs, ees, rto, support, opening, retail, price, epsa, conference, committee, message, government, intended, reply, believe, short	1276
2	order, steve, national, yesterday, presentation, commission, meeting, energy, visit, pay, regulatory, daily, policy, president, schedule, hearing, enron, affairs, thank, senate, bush, house, mara, ees, power, short, crisis, utility, position, customer, governor, communications, susan, mail, robert, gas, process, capacity, group, development, government, support, later, employees, companies	858
3	memo, attached, enron, original, ferc, california, sent, message, rto, http, proposed, draft, october, gas, peter, letter, budget, proposal, report, dernehl, lisa, yoho, styles, transmission, costs, contracts, shapiro, decision, contract, ginger, richard, european, natural, html, cost, center, enronxgate, org, language, legislation, london, chris, points, reports, development, press, aleck, opening, dadson, governor, release, president, sue, nord, communications, affairs, dynegy, action, susan, conference, authority, net, eric, market, lay, mail, ees, administration, used, public, outside, information, greg, trade, jose, project, million, capacity, meeting, rate, frank, legal, robert, competition, document, management, work, corporate, paper, rules	2610

Tabla 5.3: Palabras clave de cada área del conjunto II

ÁREA	PALABRAS CLAVE CONJUNTO III	CORREOS ASOCIA-DOS
0	bible, holiday, emazing, corp, daily, blessings, schedule, reminder, win, mailbits, abuse, tip, dollars, october, verse, meeting, internet, image, health, tuesday, aol, kevin, following, power, operations, data, scheduled, ees, netscape, jennifer, cutaia, llgm, received, neville, michelle, december	1079
1	joe, corp, llgm, update, elizabeth, saturday, brenna, days, judy, smith, mcnichols, services, company, energy, sandra, juan, daily, emazing, image, tip, internet, site, help, program, access, content, trading, market	600
2	sandoval, questions, change, loneta, report, access, following, power, ees, edison, elizabeth, angela, guillen, andrea, barnett, bible, verse, emazing, person, group, mark, deal, lsharis, juan, daily, laura, did, god, regina, hotmail, flash, tip, sent, message, original, health	744
3	image, list, jeff, yahoo, sandoval, mary, prayer, daily, blessings, nov, day, cool, coulter, kayne, click, ees, halliburton, houston, new, deals, llgm, center, fun, provide, johnson, market, power, alex, jpg, juan, aol, http, access, web, company	842

Tabla 5.4: Palabras clave de cada área del conjunto III

Conjunto	Carpetas	Archivos/# de Correos	Nodos
Conjunto I	1	160	448
Conjunto II	116	6071	3174
Conjunto III	16	3265	1561

Tabla 5.5: Descripción de los conjuntos de datos utilizados en las pruebas de identificación de temas y áreas.

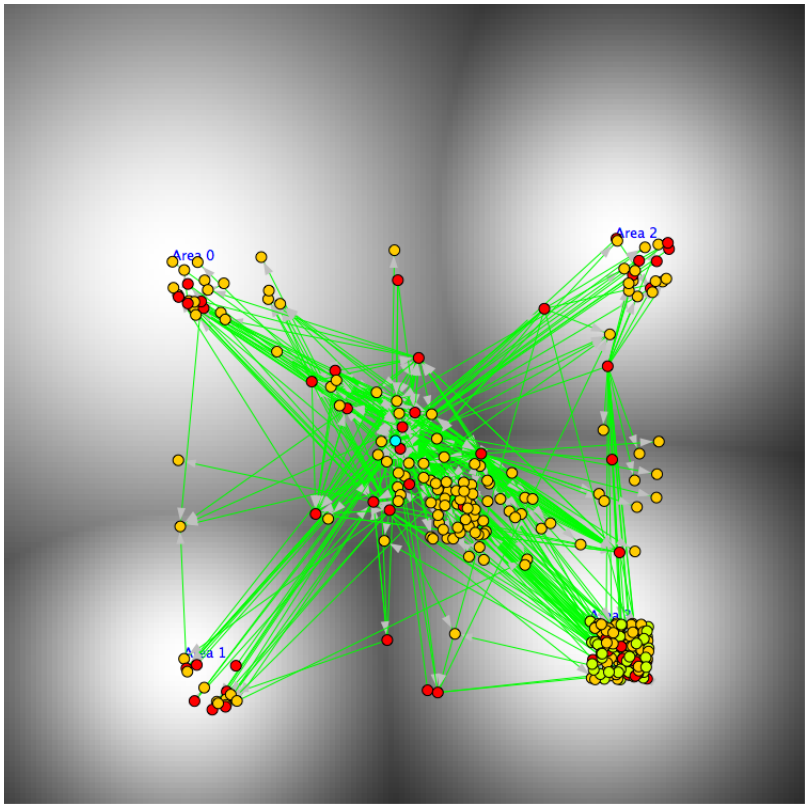
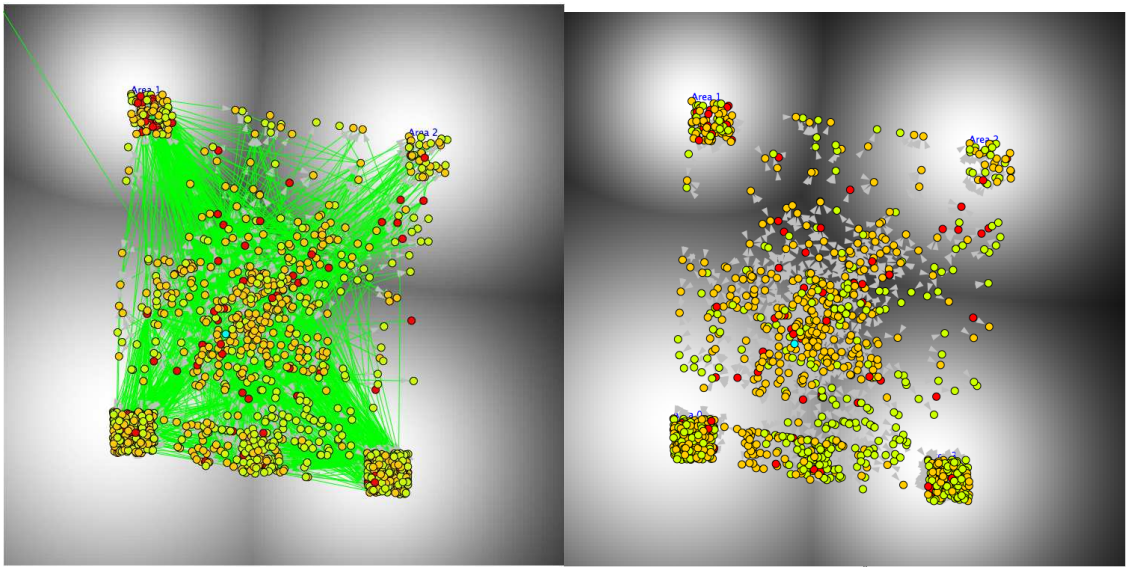


Figura 5.9: Resultado del algoritmo de pintado basado en campos de fuerza para el conjunto de datos I



(a) Ejemplo de la unificación de métodos de visualización para el conjunto III mostrando todos los arcos

(b) Ejemplo de unificación de los métodos de visualización para el conjunto III sin mostrar los arcos

Figura 5.10: Comparación entre el uso y el no uso de arcos para el resultado del uso unificado de los métodos de visualización para el conjunto III

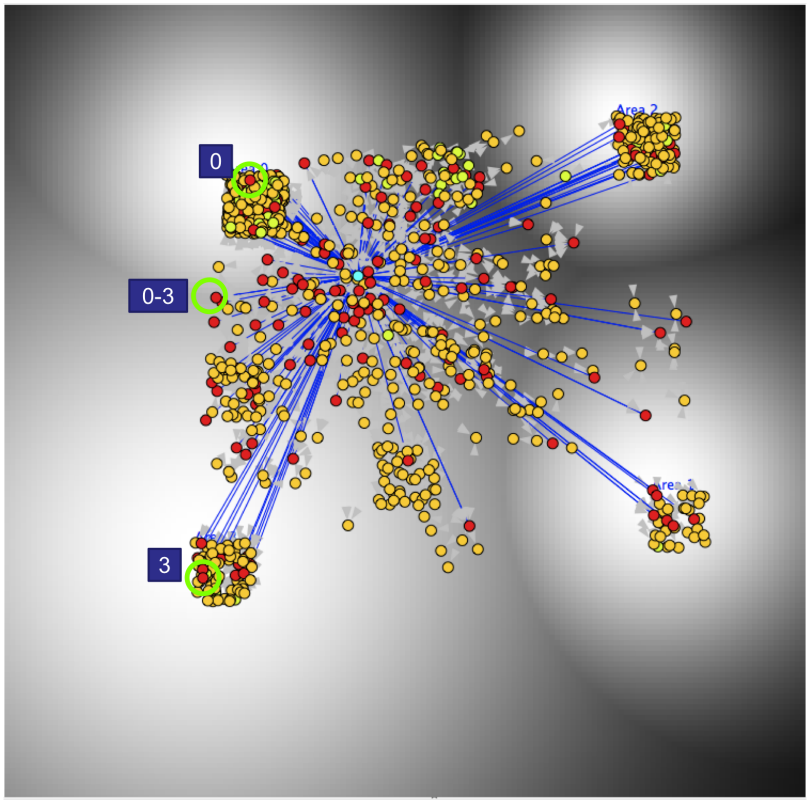


Figura 5.11: Resultado del algoritmo de pintado basado en campos de fuerza para el conjunto de datos II. Solo se presentan los arcos relacionados con el nodo *ego*

Campo	**Valor**
Área	Área 0
From	sue.nord@enron.com
To	joseph.alamo@enron.com, daniel.allegretti@enron.com, joe.allen@enron.com (y otros)
Cc	
Asunto	Government Affairs Retainer Agreements
Cuerpo	Please send out Notices of Termination by November 30, 2001 for all retainer (fixed payments) agreements (lobbying, legal, consulting, etc...) We need to end all payments as of December 31, 2001. Thank you very much.

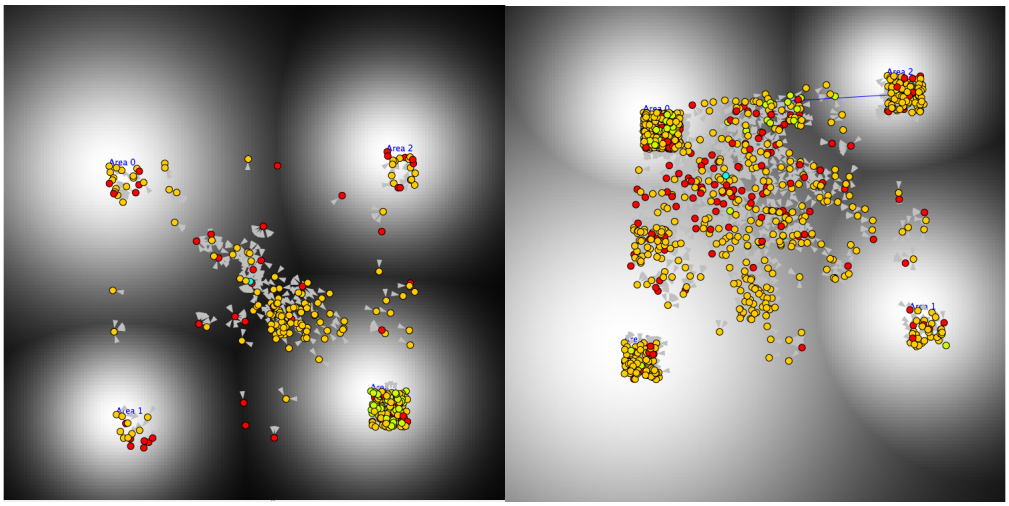
Tabla 5.6: Resumen de información del nodo de Marisa Rapaciolli

Campo	Valor
Área	Área 3
From	jonathan.whitehead@enron.com
To	steven.kean@enron.com, richard.shapiro@enron.com, tom.briggs@enron.com
Cc	
Asunto	Jose LNG Background
Cuerpo	Please find attached a brief outline of the Jose LNG project. Eric will call later this afternoon to discuss.

Tabla 5.7: Resumen de información del nodo de Jonathan Whitehead

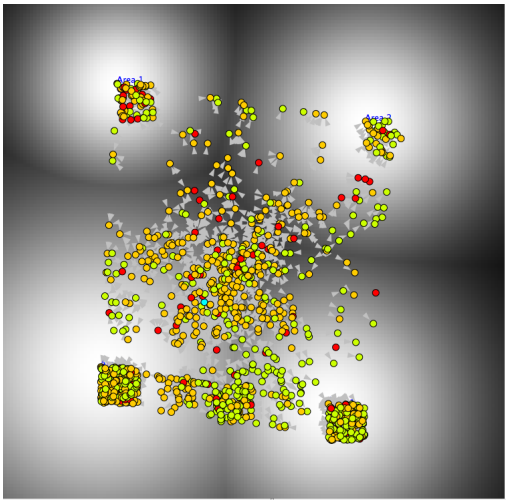
Correo	Asunto	Cuerpo	Área
1	NY RTO Order	Attached is a summary of the New York RTO Order issued by FERC late yesterday. I apologize for the length, but I wanted to include a number of quotes from the Order that repeat arguments that we have been making over the past few years. In a procedural Order issued today, FERC has scheduled the Northeast mediation process to begin next Thursday, July 19th.	0
2	Major Unbundling Victory	The NYPSC today issued an Order directing the expedited consideration of "bottoms-up" rate unbundling by all of the state's major electric and gas utilities. Enron has long pushed the PSC to address this issue. The PSC wants unbundled rates in effect by the first half of 2002, and has set up a two-stage process to get there. Under the first stage, which the PSC wants completed and brought to them for decision by this August, major policy issues of statewide import are to be considered. Resolution of these issues will guide the actual embedded cost studies which each of the utilities are directed to perform. These cost studies and the fully unbundled rates that will flow from them, will then be reviewed in utility-specific cases. The PSC expects the studies to be completed before the end of this year. I anticipate that we will actively participate in this proceeding, which is actually being spun off from the "end-state" proceeding, through our NESPA coalition. A copy of the PSC's order should be available on their website: www.dps.state.ny.us	3
3	NYPSC approves market-based backout credit for RG&E; Adopts Enron Online as Index	Another big victory for Enron in New York! In an Order issued by the NYPSC on March 29, the NYPSC has approved a proposal to modify the existing fixed backout credit of 2.8 cents per kwh for energy and capacity and replace it with a market-based backout credit. The additional retail adder of 0.4 cents per kwh would remain. Load-serving entities (LSEs) will have the option of continuing to buy their energy and capacity from RG&E at the market based backout credit which will be determined 15 days prior to the beginning of each of the two six-month capability periods, and will be based on "a forecast of energy prices derived from Enron Corporation's (Enron) EnronOnline trading platform." To my knowledge, this is the first occasion that a regulatory body has agreed to use Enron Online as an index! LSEs can elect instead to procure their own commodity and get a market-based backout credit (MBBC). Under this option, the LSEs will get a market based backout credit. However, the Enron online prices for the six months determined as above will be reconciled on a lag basis against actual market prices. The PSC has capped the amount of the reconciliation for any LSE to the product of the LSE's total load for the calendar year and $0.065 per kwh (PSC staff's estimate of RG&E's embedded power supply cost). (The PSC Order actually says $0.015 but this appears to be a mistake since the attached settlement states $0.065.) A copy of the PSC's Order is available on its website: www.dps.state.ny.us	3

Tabla 5.8: Resumen de información del nodo de Howard Fromer. Este nodo se encuentra entre las áreas 0 y 3 del conjunto.



(a) Unificación de la red de contactos y la distribución de temas para el conjunto I

(b) Unificación de la red de contactos y la distribución de temas para el conjunto II



(c) Unificación de la red de contactos y la distribución de temas para el conjunto III

Figura 5.12: Áreas y redes de contactos encontradas para los conjuntos de datos I, II y III. Se muestra la unificación de áreas de temas y las redes de contactos.

Algoritmo	Conjunto I	Conjunto II
Eades	12197 ms	30762 ms
Fruchterman & Reingold	8737 ms	49761 ms
Fuerza basado en anillos	˜12197 ms	˜30762 ms
Campo de Fuerzas	<55 ms	<100 ms
Coloreado de Nodos	389 ms	1907 ms

Tabla 5.9: Resumen de tiempos de ejecución de algoritmos de visualización de grafos

Capítulo 6

Conclusiones y Trabajo Futuro

> "... you can't connect the dots looking forward; you can only connect them looking backwards ..." – Steve Jobs' Commencement Speech, 12 June, 2005.

A través del desarrollo de este trabajo se exploraron temas relacionados con la identificación, extracción y visualización de información contenida en un conjunto de datos, que en este caso, son correos electrónicos. Esta exploración permitió identificar diversas técnicas y procedimientos para cada una de las áreas anteriormente enumeradas, pero que hasta ahora no tenían un punto de convergencia común que le diera una solución de valor agregado a los usuarios en general.

Luego de dicha exploración, se determinó que era necesario desarrollar algunas técnicas, y complementar otras, para presentar de una forma intuitiva para los usuarios la información contenida en un conjunto de correos electrónicos, que a simple vista, no tienen ninguna relación particular unos con otros, pero que, cuando son vistos como un todo, es posible identificar relaciones a dos niveles: personas y temas.

Dichos niveles, aunque no son excluyentes, no son fácilmente relacionables entre sí, y ese problema, el de encontrar un método de relacionarlos fue el objetivo de este trabajo.

Como se mencionó, para facilitar la visualización se desarrollaron algunas técnicas y se adaptaron otras:

- Coloreado de nodos: esta técnica permite asignar a cada nodo un valor de grado de vecindad, el cual es un número entero, de modo que a cada uno se le asigne un color según dicho número.

- Algoritmo de pintado por anillos: esta técnica permite agrupar los nodos de una red de contactos según el grado de cercanía que guarden con el nodo que representa al dueño de los mensajes de correo. Adicionalmente, asume que no existe una alta conectividad entre nodos, lo que hace que no pueda ser aplicado a cualquier grafo.

- Identificación de áreas temáticas: la identificación de áreas temáticas permite agrupar los mensajes de correo electrónico por temas, es decir, asignar una etiqueta a un subconjunto de los mensajes de correo electrónico que tienen cierta similitud. La similitud está dada por la cercanía temática entre ellos.
  La identificación de las áreas permite además, definir espacios geométricos en los cuales serán asignados los nodos de la red de contactos posteriormente.

- Algoritmo basado en campos de fuerza: este algoritmo utiliza la relación que tiene cada uno de los nodos con los correos electrónicos del conjunto estudiado y posteriormente utiliza la información de las áreas temáticas identificadas para asignarle una posición dentro del mapa de áreas a cada uno de los nodos.

La adaptación y posterior unificación de las técnicas permitió encontrar un método de visualización intuitivo de la información sobre la información contenida en un conjunto de correos electrónicos, cumpliendo el objetivo propuesto para este trabajo.

Todos los prototipos desarrollados durante este trabajo han sido pensados para ser ejecutados como un proceso ejecutado en lote, en donde se parte de un conjunto estable de mensajes de correo electrónico. Esto hace que el análisis de los temas y la creación de la red de contactos sea, sencilla pero demorada, razón por la cual se plantea el desarrollo de un esquema de identificación de temas y visualización de áreas incremental.

En el caso de las redes sociales, debido a la forma en que se definió la ecuación para la construcción de las redes, es posible partir de un solo correo electrónico y modificar la red a medida que se van agregando nuevos correos y actores.

Este nuevo esquema de identificación de temas, construcción de áreas y asociación con la red de contactos, puede permitir además identificar la evolución de los temas y de la aparición de los mismos en el mapa de las áreas.

6.1. Trabajo Futuro

El trabajo relacionado con extracción de información y descubrimiento de conocimiento en texto tiene muchos campos de aplicación y sigue siendo un área de estudio amplia. En el caso específico de este trabajo, el descubrimiento y visualización de información y conocimiento no evidente se tiene la misma condición.

Dentro de los posibles trabajos futuros relacionados con este trabajo están los relacionados con el desempeño del pre-procesamiento y organización de los datos, mecanismos de identificación y pintado de áreas, procesamiento de información en tiempo real y vinculación con clientes de correo.

Algunos tópicos a tratar como trabajo futuro incluyen pero no se limitan a:

- Uno de los pasos más demorados del método planteado es el pre-procesamiento de los datos, específicamente la búsqueda del número óptimo de áreas de temas. Esto es porque hasta ahora se hace procesando miles de mensajes de correo y de palabras en lote. La idea es mejorar el proceso para realizarlo en tiempo real, para lo cual es posible utilizar métodos como Growing SOM o algún método adaptativo que permita agregar nueva información de temas y su clasificación a medida que van llegando los mensajes, que es de hecho, lo que pasa en el día a día de las personas al recibir los mensajes personales o del trabajo.

- Identificación del contexto de los mensajes. Un trabajo interesante puede resultar al buscar un método que permita identificar, o por lo menos aproximar, el contexto en el cual han sido desarrollados los mensajes de correo de un usuario. Esto puede ser posible si se utiliza la información histórica de los mensajes ya procesados y se utiliza algún método como Análisis Formal de Conceptos.

- Una vez se haya desarrollado el esquema de evaluación incremental, es posible convertir el desarrollo en un *add-on* para algún cliente de correo electrónico como Thunderbird.

- Utilización de los métodos y resultados obtenidos para construir una herramienta de gestión de conocimiento que permita identificar expertos en diferentes temas a lo largo de una organización. Adicionalmente, que permita identificar buenas prácticas de proyectos corporativos y transferirlas a otros equipos en la organización.

Bibliografía

[1] S. Wasserman and K. Faust, *Social Network Analysis: Methods and Applications*. Cambridge Press, Nov 25 1994.

[2] T. M. J. Fruchterman and E. M. Reingold, "Graph drawing by force-directed placement," *Software - Practice and Experience*, vol. 21, no. 11, pp. 1129–1164, 1991.

[3] T. F. Cox and M. A. A. Cox, *Multidimensional Scaling*, vol. 1 of *Monographs on statistics and applied probaility*. Chapman & Hall/CRC, 2000.

[4] D. Boyd, "Faceted id/entity: Managing representation in a digital world," Master's thesis, The Massachusetts Institute of Technology, the MIT Media Lab, September 2002.

[5] D. W. McDonald, "Recommending collaboration with social networks: a comparative evaluation," in *CHI '03: Proceedings of the SIGCHI conference on Human factors in computing systems*, (New York, NY, USA), pp. 593–600, ACM Press, 2003.

[6] M. H. Zack, "Researching organizational systems using social network analysis," in *HICSS '00: Proceedings of the 33rd Hawaii International Conference on System Sciences-Volume 7*, (Washington, DC, USA), p. 7043, IEEE Computer Society, 2000.

[7] V. R. Carvalho, W. Wu, and W. W. Cohen, "Discovering leadership roles in email workgorups," in *Proceedings of Fourth Conference on Email and Anti-Spam – CEAS 2007*, (Mountain View, CA), August 2 – 3 2007.

[8] L. Johansen, M. Rowell, K. Butler, and P. McDaniel, "Email comunities of interest," in *Proceedings of Fourth Conference on Email and Anti-Spam, CEAS 2007*, (Mountain View, CA), August 2 – 3 2007.

[9] C. Neustaedter, A. B. Brush, and M. A. Smith, "The social network and relationship finder: Social sorting for email triage," in *Proceedings of Second Conference on Email and Anti-Spam CEAS 2005*, (Stanford University), July 21 & 22 2005.

[10] J. Golbeck and J. Hendler, "Reputation network analysis for email filtering," in *Proceedings of First Conference on Email and Anti-Spam (CEAS)*, (Mountain View, CA), July 30 – 31 2004.

[11] R. Khoussainov and N. Kushmerick, "Email task management: An iterative relational learning approach," in *Proceedings of Second Conference on Email and Anti-Spam CEAS 2005*, (Stanford University), July 21 & 22 2005.

[12] M. A. Smith, J. Ubois, and B. M. Gross, "Forward thinking," in *Proceedings of Second Conference on Email and Anti-Spam CEAS 2005*, (Stanford University), July 21 & 22 2005.

[13] A. C. Surendran, J. C. Platt, and E. Renshaw, "Automatic discovery of personal topics to organize email," in *Proceedings of Second Conference on Email and Anti-Spam CEAS 2005*, (Stanford University), July 21 & 22 2005.

[14] A. Secker, A. Freitas, and J. Timmis, "Aisec: An artificial immune system for e-mail classification," in *Proceedings of the Congress on Evolutionary Computation* (R. Sarker, R. Reynolds, H. Abbass, T. Kay-Chen, R. McKay, D. Essam, and T. Gedeon, eds.), (Canberra. Australia), pp. 131–139, IEEE, December 2003.

[15] E. Minkov and W. W. Cohen, "An email and meeting assistant using graph walks," in *Proceedings of Third Conference on Email and Anti-Spam CEAS 2006*, (Mountain View, CA), July 27 – 28 2006.

[16] P. Domingos and M. Richardson, "Mining the network value of customers," in *KDD '01: Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, (New York, NY, USA), pp. 57–66, ACM Press, 2001.

[17] S. Chakrabarti, M. M. Joshi, K. Punera, and D. M. Pennock, "The structure of broad topics on the web," in *WWW '02: Proceedings of the 11th international conference on World Wide Web*, (New York, NY, USA), pp. 251–262, ACM Press, 2002.

[18] B. Amento, L. Terveen, W. Hill, D. Hix, and R. Schulman, "Experiments in social data mining: The topicshop system," *ACM Trans. Comput.-Hum. Interact.*, vol. 10, no. 1, pp. 54–85, 2003.

[19] M. K. Ahuja, D. F. Galletta, and K. M. Carley, "Individual centrality and performance in virtual r&d groups: An empirical study," *Manage. Sci.*, vol. 49, no. 1, pp. 21–38, 2003.

[20] D. Liben-Nowell and J. Kleinberg, "The link prediction problem for social networks," in *CIKM '03: Proceedings of the twelfth international conference on Information and knowledge management*, (New York, NY, USA), pp. 556–559, ACM Press, 2003.

[21] B. J. Mirza, B. J. Keller, and N. Ramakrishnan, "Studying recommendation algorithms by graph analysis," *J. Intell. Inf. Syst.*, vol. 20, no. 2, pp. 131–160, 2003.

[22] A. J. O'Donnell, W. C. Mankowski, and J. Abrahamson, "Using email social network analysis for detecting unauthorized accounts," in *Proceedings of Third Conference on Email and Anti-Spam CEAS 2006*, (Mountain View, CA), July 27 – 28 2006.

[23] R. Cross and A. Parker, *The Hidden Power of Social Network: Understanding How Work Really Gets Done in Organizations*. Harvard Business School Press, 2004.

[24] D. Kusui, K. Tateishi, and T. Fukushima, "Information extraction and visualization from internet documents," *NEC Journal of Advanced Technology*, vol. 2, pp. 157 – 163, Spring 2005.

[25] S. Chakrabarti, "Data mining for hypertext: a tutorial survey," *SIGKDD Explor. Newsl.*, vol. 1, no. 2, pp. 1–11, 2000.

[26] J. A. Black and N. Ranjan, "Automated event extraction from email." Final Report of CS224N/Ling237 Course in Stanford: http://nlp.stanford.edu/courses/cs224n/2004/, Spring 2004.

[27] I. L. MacDonald and W. Zucchini, *Hidden Markov and Other Models for Discrete-Valued Time Series*. Chapman & Hall/CRC, 2002.

[28] L. Denoyer, H. Zaragoza, and P. Gallinari, "Hmm-based passage models for document classification and ranking," in *Proceedings of 23rd BCS European Annual Colloquium on Information Retrieval*, 2001.

[29] S. Sarawagi and W. W. Cohen, "Semi-markov conditional random fields for information extraction," in *Advances in Neural Information Processing Systems*, December 13 2004.

[30] J. Hertz, A. Krogh, and R. G. Palmer, *Introduction to the Theory of Neural Computation*. Santafe Institute, 1999.

[31] P. Ojha, "Neural networks and decision trees." Knowledge Based Systems Course Notes at University of Ulster, 2003. Disponible en: http://www.infj.ulst.ac.uk/ cbdq23/teaching/com812j/notes/autokacq.pdf.

[32] M. Berthold and D. J. hand, *Intelligent Data Analysis: an Introduction*. Springer, 1999.

[33] S. Forrest, A. Perelson, L. Allen, and R. Cherukuri, "Self-nonself discrimination in a computer," in *Proceedings of the 1994 IEEE Symposium on Research in Security and Privacy*, (Oakland, CA), pp. 202–212, IEEE Computer Society Press, May 16 – 18 1994.

[34] J. Kim and P. Bentley, "The artificial immune system for network intrusion detection: An investigation of clonal selection with a negative selection operator," 2001.

[35] J. Twycross and S. Cayzer, "An immune-based approach to document classification," Tech. Rep. HPL-2002-292, Information Infrastructure Laboratory HP Laboratories Bristol, October 30 2002.

[36] G. Marwah and L. Boggess, "Artificial immune systems for classification : Some issues," 2002.

[37] F. Picarougne, G. Venturini, and C. Guinot, "Un algorithme génétique parallèle pour la veille stratégique sur internet," in *Veille Stratégique Scientifique et Technologique VSST 2004*, vol. 2, (Toulouse, France), pp. 519 – 528, Octobre 25 – 29 2004.

[38] M. Sahami, S. Dumais, D. Heckerman, and E. Horvitz, "A bayesian approach to filtering junk E-mail," in *Learning for Text Categorization: Papers from the 1998 Workshop*, (Madison, Wisconsin), AAAI Technical Report WS-98-05, 1998.

[39] B. Rhodes, J. Mahaffey, and J. Cannady, "Multiple self-organizing maps for intrusion detection," in *National Information Systems Security Conference*, 2000.

[40] F. Fernández, "Redes neuronales artificiales no supervisadas." Departamento de Informática Universidad Carlos III de Madrid, Marzo 2004.

[41] F. González and D. Dasgupta, "Anomaly detection using real-valued negative selection." Genetic Programming and Evolvable Machines, 2003.

[42] U. Brandes and T. Willhalm, "Visualization of bibliographic networks with a reshaped landscape metaphor," in *VISSYM '02: Proceedings of the symposium on Data Visualisation 2002*, (Aire-la-Ville, Switzerland, Switzerland), pp. 159–ff, Eurographics Association, 2002.

[43] D. Masterson and N. Kushmerick, "Information extraction from multi-document threads," in *Proceedings of the International Workshop on Adaptive Text Extraction and Mining*, (Catvat – Dubrovnik (Croatia)), September 22 2003.

[44] U. Priss, "Formal concept analysis in information science," *Annual Review of Information Science and Technology, ARIST*, vol. 40, pp. 521 – 543, 2006.

[45] U. Priss, "A graphical interface for document retrieval based on formal concept analysis," in *Proceedings of the 8th Midwest Artificial Intelligence and Cognitive Science Conference* (E. Santos, ed.), (Dayton Ohio), pp. 66 – 70, Wright State University, May 30 – June 1 1997.

[46] P. Cimiano, A. Hotho, G. Stumme, and J. Tane, "Conceptual knowledge processing with formal concept analysis and ontologies," in *Proceedings of the The Second International Conference on Formal Concept Analysis (ICFCA 04)* (Springer, ed.), pp. 189 – 207, 2004.

[47] I. Richard J. Cole and P. W. Eklund, "Analyzing an email collection using formal concept analysis," in *PKDD '99: Proceedings of the Third European Conference on Principles of Data Mining and Knowledge Discovery*, (London, UK), pp. 309–315, Springer-Verlag, 1999.

[48] E. M. Awad and H. M. Ghaziri, *Knowledge Management*. Prentice Hall, 1st ed., 2003.

[49] I. Nonaka and H. Tackeuchi, *The knowledge creating company*. Oxford University Press, 1995.

[50] P. Busch, D. Richards, and C. N. G. K. Dampney, "The graphical interpretation of plausible tacit knowledge flows," in *CRPITS '24: Proceedings of the Australian symposium on Information visualisation*, (Darlinghurst, Australia, Australia), pp. 37–46, Australian Computer Society, Inc., 2003.

[51] P. A. Busch, D. Richards, and C. N. G. K. Dampney, "Visual mapping of articulable tacit knowledge," in *CRPITS '01: Australian symposium on Information visualisation*, (Darlinghurst, Australia, Australia), pp. 37–47, Australian Computer Society, Inc., 2001.

[52] P. A. Eades, "A heuristic for graph drawing.," in *Congressus Numerantium*, vol. 42, pp. 149–160, 1984.

[53] L. C. Freeman, "Graphics techniques for exploring social network data," in *Models and Methods in Social Network Analysis*, Peter J. Carrington and John Scott and Stanley Wasseman, 2005.

[54] B. Klimt and Y. Yang, "Introducing the enron corpus," in *Proceedings of Confenrence on Email and Anti Spam, CEAS*, 2004.

[55] H. C. Purchase, J.-A. Allder, and D. Carrington, "Graph layout aesthetics in uml diagrams : User preferences," *Journal of Graph Algortihms an Applications*, vol. 6, no. 3, pp. 255 – 279, 2002.

[56] M. Kantardzic, *Data Mining: Concepts, Models, Methods, and Algorithms*. IEEE Press, 2001.

[57] O. Gospodnetic and E. Hatcher, *Lucene in Action*. Manning Publications, 2004. GOS o 05:1 1.Ex.

[58] M. Porter, V. Rijsbergen, and S. Robertson, "New models in probabilistic information retrieval," tech. rep., British Library Research and Development Report, no. 5587, 1980.

[59] A. Romero and F. Nino, "Keyword extraction using an artificial immune system," in *GECCO '07: Proceedings of the 9th annual conference on Genetic and evolutionary computation*, (New York, NY, USA), pp. 181–181, ACM, 2007.

[60] P.-N. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining*. Addison – Wesley, 2005.