Gestión de Conocimiento a través del Análisis de las Redes Sociales y la Información Contenida en los Correos Electrónicos Personales

Juan David Cruz Gómez, 299618
Seminario de Investigación I
Maestría en Ingeniería – Ingeniería de Sistemas y Computación

***Abstract*— Debido al gran interés mostrado por diferentes organizaciones en la gestión de su capital intelectual, se han desarrollado diferentes métodos y herramientas para llevar a cabo esta tarea, la cual requiere la integración de diferentes soluciones. Este artículo presenta algunas de las herramientas utilizadas para construir soluciones de gestión de conocimiento tácito y explícito así como algunas de sus aplicaciones buscando un no un enfoque corporativo sino personal.**

***Index Terms*— Análisis de redes sociales, gestión de conocimiento, trabajo colaborativo, clasificación de documentos, gestión de correo**

I. Introducción

La gestión del capital intelectual ha despertado el interés de las grandes compañías por buscar y desarrollar herramientas que le permitan gestionar dicho capital de la mejor forma posible. Para esto, se ha desarrollado la gestión de conocimiento y dentro de esta, diferentes conceptos como los diferentes tipos de conocimiento y herramientas para el manejo adecuado de dichos tipos, como el análisis de redes sociales. Sin embargo, estos conceptos han trascendido fuera de de las empresas para se ahora parte de la investigación sobre como hacer el día a día de las personas más fácil, o al menos una parte. De esta forma, se han desarrollado técnicas basadas en dichas teorías para clasificar y decidir que hacer con los correos electrónicos de una persona de forma automática; pero, adicionalmente a la tarea de clasificación, se pueden añadir otras ayudas como la extracción de conocimiento de los textos de los correos y utilizar dicho conocimiento para potenciar su uso.

Las redes sociales son una poderosa herramienta para analizar las relaciones entre personas y/u organizaciones. Actualmente, son utilizadas para gestionar el conocimiento, descubrir expertos y recomendar colaboraciones entre personas ya que contienen el flujo de la información. Sin embargo, aunque las redes sociales ya son una herramienta bastante útil, la información que aportan puede ser potenciada si se le añade la información sobre los documentos y sus relaciones que un individuo posee sobre diferentes temas.

El trabajo está organizado de la siguiente manera: en la sección II se define el problema y se explica una taxonomía de técnicas que ayudan a resolverlo, en la sección III, se explica el concepto de gestión de conocimiento, que busca y cuales son los tipos de conocimiento, en la sección IV se aborda el tema de las redes sociales desde su definición hasta el análisis de las mismas explicando algunas de las medidas más utilizadas para esto, así como algunas de sus aplicaciones, en la sección V se explican algunas técnicas relevantes para la extracción de conocimiento y la clasificación de documentos, en la sección VI se establecen algunas conclusiones y en la última sección se plantean algunas preguntas de investigación que surgen de este trabajo.

II. Definición del problema

Los medios de comunicación han evolucionado, ya no se tienen únicamente aquellos que son impresos, como cartas o libros, ni aquellos electrónicos que solo proveen información, que no permiten una interacción real y que permitan el crecimiento intelectual de las partes involucradas. El vertiginoso crecimiento de las tecnologías de información y comunicaciones, apoyado en el desarrollo de Internet, han permitido la aparición de nuevos canales de comunicación, que permiten la interacción de personas, haciendo que el conocimiento que cada una de dichas personas posee sea exteriorizado y se comiencen a construir diferentes bases de conocimiento, que entrega bases para generar más conocimiento en las personas.

Uno de los canales más importantes, y, más utilizados hoy en día, es el correo electrónico. Una de las principales ventajas que posee es que el intercambio de información se puede hacer casi en tiempo real, logrando que en principio un intercambio seguido de mensajes entre dos personas sobre un tema se parezca a una conversación, a bajo costo, escrita y oficial, lo que ha hecho que el volumen diario de correos electrónicos aumente de forma casi exponencial[1].

Este alto volumen de información permite tener bases de datos de gran tamaño, con un gran contenido de conocimiento explícito y tácito. Adicionalmente, las personas producen y descargan documentos sobre algunos temas de su interés, generando una base de datos documental, la cual, generalmente no se relaciona de ninguna

[1]En 2005 se espera que existan 1.2 billones de cuentas de correo y el número de mensajes enviados sea superior a 36 billones en todo el mundo. *Fuente: http://archives.cnn.com/TECH/*

forma con la primera. En la figura 1 se puede ver el esquema descrito.

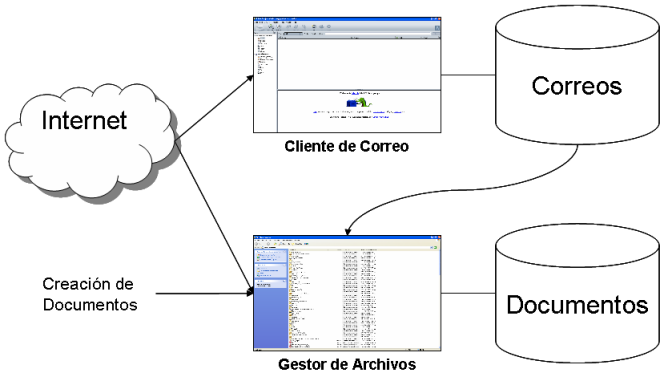

Fig. 1. Manejo general de información personal. El problema con este esquema es que no existe una relación entre los correos y los documentos personales.

La idea entonces es lograr que se puedan relacionar las dos bases de datos, que son conocimiento explícito y generar redes sociales a partir de los mensajes de correo entrantes y salientes almacenados en la base de datos. Estas redes sociales, con las relaciones de documentos y correos, van a permitir gestionar el conocimiento tácito, logrando de esta forma potenciar toda la información disponible. En la figura 2 se puede ver un esquema de lo explicado anteriormente.

La idea es entonces diseñar un sistema que esté en capacidad para crear mapas conceptuales de documentos realizando un análisis de los textos contenidos en estos, generar redes sociales de conocimiento a partir de los mensajes de correos electrónicos y relacionar esta información con el fin de obtener la mayor cantidad de conocimiento de la información almacenada por aparte.

Para realizar esta tarea se deben realizar algunas acciones con los documentos y los correos. Es necesario realizar un análisis de los documentos y los correos para relacionarlos, extraer la información para generar las redes sociales y utilizar las medidas usuales (sección IV-B) para gestionar el conocimiento tácito contenido en ellas. En las secciones siguientes se explican diferentes métodos para llevar a cabo estas tareas.

II-A. Taxonomía de una Solución de Gestión de Conocimiento

El problema descrito anteriormente requiere la integración de diferentes soluciones, de modo que se puedan gestionar e integrar los diferentes conocimientos. Para esto se propone la taxonomía mostrada en la figura 3.

El problema planteado está compuesto principalmente, de tres tópicos generales: gestión de conocimiento, análisis de redes sociales y extracción de conocimiento. El primer tópico es con el que realiza la integración de los dos tipos de conocimiento. Con el análisis de redes sociales y la extracción de conocimiento se busca apoyar la gestión de conocimiento tácito y explícito respectivamente.

III. Gestión de Conocimiento

La gestión del conocimiento tiene variadas definiciones, depende del punto de vista desde donde se den. En [1] se pueden ver algunas de ellas. Sin embargo, en general se puede decir que es la gestión del capital intelectual en una organización, con la finalidad de agregar un valor a los productos y servicios ofrecidos por dicha organización así como diferenciarlos competitivamente.

El conocimiento surge a partir de la información, que a su vez surgió de la transformación de datos obtenidos de la realidad. Es decir, a partir de la realidad se obtienen una serie de datos a partir de atributos o representaciones de dicha realidad, estos son procesados a través de cálculos, ordenamientos para convertirlos en información, la cual será utilizada para ejecutar tareas de clasificación, calificación, cuantificación, etc., para finalmente obtener el conocimiento, a partir del cuál se puede inferir o juzgar, en general, tener sabiduría [1].

El conocimiento, puede ser dividido en dos grupos: el conocimiento tácito, que se refiere a todo el conocimiento que poseen las personas, que no se puede extraer fácilmente y que no está presente en las bases de datos de las empresas, y el conocimiento explícito, el cual se compone de reglas, procedimientos, documentos. A continuación se explican con más detalle cada uno de ellos.

III-A. Conocimiento tácito

El conocimiento tácito o implícito es aquel que poseen las personas y que no puede ser fácilmente extraido. Es en general todo aquel conocimiento que no tiene una única regla para representarlo. El ejemplo más claro es conocimiento que una persona tiene para montar en bicicleta: una persona puede saber como montar en bicicleta, pero si le explica el método a otra que no sabe como montar, seguramente no va a prender y se va a caer.

Este es uno de los principales problemas en la gestión de conocimiento, como capturar correctamente y gestionar el conocimiento tácito; por ejemplo en [2] 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 (ver sección IV). Dada la dificultad de trabajar con este tipo de conocimiento, en [3] 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.

III-A.1. 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.

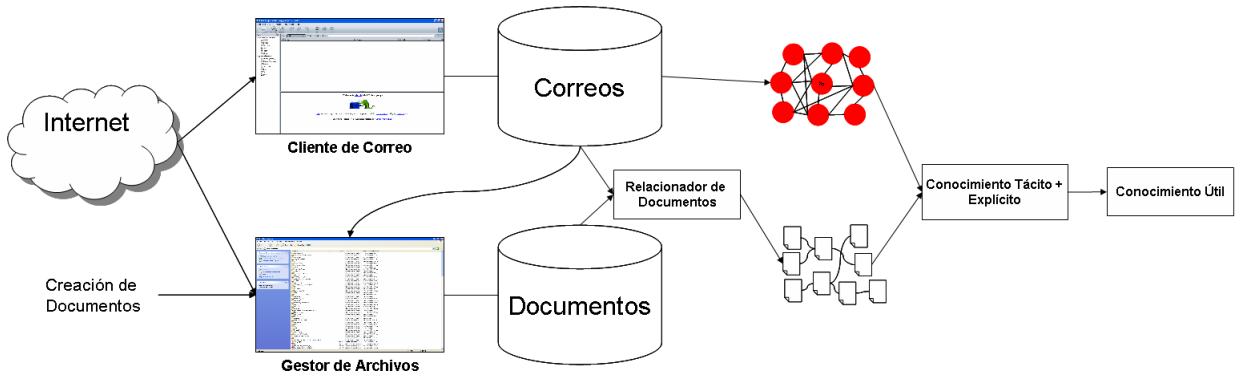

Fig. 2. Manejo propuesto de los documentos y correos. La idea es establecer una relación entre los documentos, conocimiento explícito y las redes sociales, que son vistas como una forma de gestionar el conocimiento tácito.

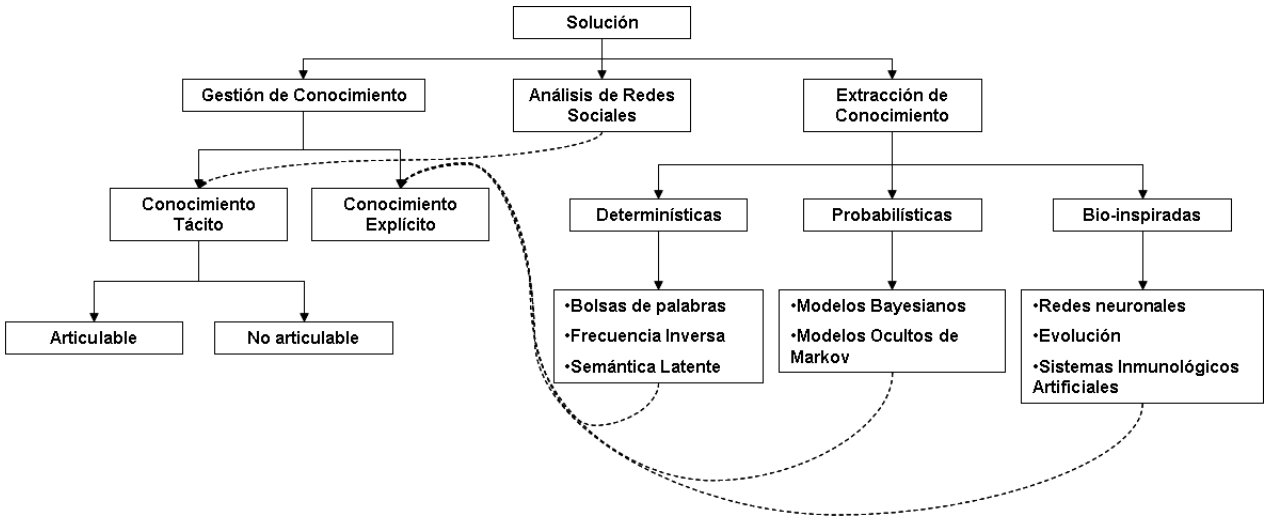

Fig. 3. Taxonomía para una solución de gestión de conocimiento. Se deben integrar diferentes soluciones para gestionar e integrar los dos tipos de conocimiento.

III-A.2. 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.

III-B. Conocimiento explícito

El conocimiento explícito es todo aquel conocimiento que está disponible en documentos, bases de datos o cual 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.

III-C. Ciclo de creación del conocimiento

Según Nonaka [4], el conocimiento dentro de una organización se genera a partir de cuatro fases:

1. Socialización: en esta fase, los empleados comparten sus experiencias e ideas con otros. De esta forma el conocimiento tácito individual se convierte en conocimiento tácito colectivo.
2. Externalización: en esta fase, el conocimiento tácito colectivo es exteriorizado de alguna forma por cada uno de los empleados de forma individual. Se supone que solo el conocimiento tácito articulable puede ser explicitado de una forma relativamente sencilla de modo que se genera conocimiento explícito.
3. Combinación: en esta fase cada una de las generaciones de conocimiento explícito es combinada, ya sea por el intercambio de documentos, correos electrónicos e informes entre otros.
4. Interiorización: en esta fase el conocimiento explícito colectivo es asimilado por cada uno de los empleados, generando así nuevo conocimiento tácito individual y permitiendo que el ciclo comience de nuevo.

En la figura 4 se puede ver un esquema de este ciclo.

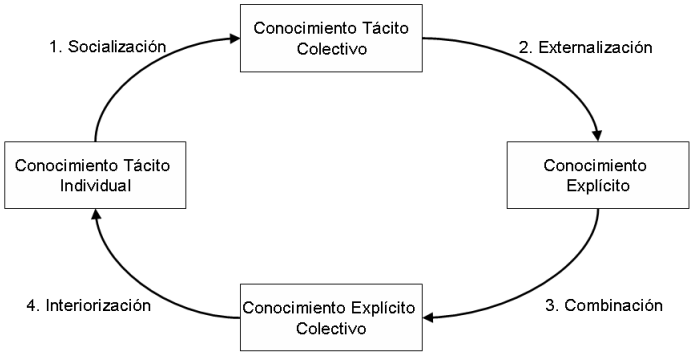

Fig. 4. Ciclo de creación del conocimiento[4].

Se debe tener en cuenta entonces, 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 conposterioridad. 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.

III-D. Algunas aplicaciones de gestión de conocimiento

La gestión de conocimiento tiene actualmente variadas aplicaciones. La más común es la implementación en organizaciones de proyectos que buscan identificar los tópicos que la harán más competitiva así como de las herramientas con lo que lo lograrán hacer tales como herramientas colaborativas.

Una aplicación enfocada a organizaciones es la propuesta por [5], en donde se busca integrar la gestión de conocimiento con la inteligencia de negocio (BI por sus iniciales en inglés). La idea principal de esta aplicación es utilizar las herramientas y teorías de gestión de conocimiento para aprovechar al máximo toda la información que proveen las herramientas de BI como los reportes, minería de datos y bases de datos multidimensionales para dar una ventaja competitiva a la organización.

Otra aplicación es la propuesta por [6], en donde se utiliza para potenciar a los sistemas multi-agentes. Esta es una aplicación mucho más académica que empresarial, sin embargo es importante ya que en el diseño de este tipo de sistemas, se parte de una organización social, por lo que resulta natural aplicarla. Otra propuesta similar es la de [7], en donde se explica un modelo general para la creación de sistemas expertos basados en ingeniería y gestión de conocimiento.

Otra aplicaciónque tiene que ver con el trabajo colaborativo, es la referente a la identificación de expertos. En [8], se propone un método para descubrir expertos a través del análisis de los mensajes de correo que se envían en una empresa. La idea es analizar el texto de los documentos con técnicas de clustering y luego crear un grafo con la información del remitente y los destinatarios. Esto lo llaman grafo de experticia, al cual le aplican algunas técnicas para descubrir los expertos en este. Esto es una red social, la diferencia es que no utilizan las técnicas de análisis de redes sociales para descubrir la información.

En [9] se propone una técnica de análisis de correos que permite agregar valor a los mensajes de las personas a través del estudio de su contenido y la búsqueda de información relacionada con el tema en Internet.

IV. Redes Sociales

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 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 más sub-redes, también es utilizada para gestionar el capital intelectual de la organización, descubrir quién sabe que.

IV-A. 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 natural 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 que 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) o 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.

Existe otra forma de generar una red social, 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.

IV-B. Análisis

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 regida por la teoría de manejo 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:
  - 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.
  - 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
  - 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 vínculo 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 bidireccional.
- Medidas generales
  - 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 toros 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 dispocisión los caminos más cortos hacia otros individuos, por lo que se encuentran en una excelente posición para monitorear el flujo de información que fluye a través de esta.

IV-C. 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.

IV-C.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 la duda o sepa resolver el problema de otra sin que esta última sepa de su existencia. 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 [10].

En [11] ven a una organización como una red actores en la que existen relaciones del tipo *quién quiere que con quienes.* Para responder esta clase de preguntas, utilizan el análisis de redes sociales. Definen 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.

IV-C.2. Gestión de correos: El correo electrónico es hoy en día una de las herramientas de comunicación más 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 recepción diaria de correos.

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

En [12] proponen 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 la propuesta por [13], 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 [14] 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.

En [15] se 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 [12].

En [16], 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 [17], 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 extraida del comportamiento del usuario. Aunque esta última aproximación no tiene en cuanta de forma explícita la red social del usuario, es una forma interesante y novedosa de gestión de mensajes.

IV-C.3. Otras aplicaciones: La aplicación para la que 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 [18], 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 [19], se 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 [20] se propone un estudio acerca de la centralidad y el desempeño de individuos en comunidades de investigación y desarrollo. La hipotesis 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.

En [21], 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 [22] se 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.

En [23] se 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.

V. Métodos de Extracción de Conocimiento de documentos

Los documentos son el resultado de diferentes procesos en empresas y ámbitos académicos. Se puede decir, que en general, los documentos aprobados para ser publicados (en conferencias o en empresas), están bien estructurados y han sido desarrollados en torno a algún tema bien definido.

Hoy en día, el volumen de información contenida en documentos, y de forma más general, textos (por ejemplo correos electrónicos) disponibles electrónicamente es muy

alto y presenta tasas de crecimiento muy grandes, lo que hace que su gestión sea complicada y se requieran herramientas sofisticadas de modo que se pueda extraer información de ellos de la forma más eficiente posible. Para esto, una de las cosas que se hace es buscar el tema al cual pertenece un texto determinado, y con esto clasificarlo, crear mapas conceptuales o decidir que acción realizar con el, entre otros.

Con el fin de extraer esta información, se han desarrollado diferentes técnicas, algunas de ellas son diseñadas 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.

V-A. Técnicas determinísticas

Estas técnicas generalmente 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 problema con esta técnica 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 *"stop list"* o una lista de palabras excluídas, 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 [24] 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) que está dada por:

$$IDF\left(t\right) = 1 + \log \frac{N_t}{N}$$

donde N_t es el número de documentos en los que el término t ocurre y N es el número de documentos y $\frac{N_t}{N}$ es una medida de la importancia del término. Entonces la medida $IDF\left(t\right)$ sirve para modificar la longitud de los ejes según la importancia del término que representan. Con esto, se puede determinar la $t-$ésima coordenada de d está dada por:

$$TFIDF = \frac{n\left(d,t\right)}{\|d\|} \times IDF\left(t\right)$$

donde $n\left(d,t\right)$ es el número de veces que ocurre el término t en el documento d, $\|d\|$ es la norma del documento y $TFIDF$ es *la frecuencia de aparición de términos inversa del documento.*

Al tratar los documentos como puntos en un espacio $t-$dimensional, es casi natural pensar en utilizar técnicas derivadas del álgebra lineal para tratarlos. A continuación se discutirán de forma breve algunas de dichas técnicas.

Una de las acciones a realizar cuando se trabaja con documentos es compararlos, es decir, medir su similitud semántica aún cuando los documentos no contengan términos iguales. Un ejemplo se da cuando un texto contiene el término *carro* y otro contiene el término *automóvil*, los cuales son diferentes, pero están relacionados entre si. Para esto existe una técnica llamada *indexamiento por semántica latente* [24] (LSI por sus iniciales en inglés) la cual, utiliza nociones de álgebra lineal de la siguiente forma:

Sea A una matriz de términos por documento, donde a_{ij} es 0 si el término i no ocurre en el documento j y 1 en otro caso. Si dos palabras diferentes, pero semánticamente relacionadas, aparecen en diferentes documentos, entonces las filas correspondientes a dichos términos tendrán cierta similitud, en general se puede decir que varias columnas y/o filas son redundantes. Ahora, sean $\sigma_1, \sigma_2, \sigma_3, ..., \sigma_r$, donde r es el rango de la matriz A, los valores simples [24] de A y donde $|\sigma_1| \geq |\sigma_2| \geq |\sigma_3| \geq ... |\sigma_r|$; luego se aplica el método de descomposición de valores simples [24] (SVD por sus iniciales en inglés), lo que factoriza la matriz A en UDV^T, donde D corresponde a $diag\left(\sigma_1, \sigma_2, \sigma_3, ..., \sigma_r\right)$ y U y V son matrices ortogonales. El método no utiliza todos los posibles valores de la matriz, sino que se utilizan los primeros k valores de A, de modo que se tiene la siguiente aproximación: $U_k D_k V_k^T$, donde k es un parámetro ajustable. Cada columna de $U_k$representa un término como un vector $k-$dimensional y cada fila de V_k representa un documento como un vector $k-$dimensional. En general, el método consiste en llevar un término a un vector $k-$dimensional y encontrar documentos que se encuentren cerca.

Uno de los principales inconvenientes de esta técnicas es la alta dimensionalidad con la que se pueden encontrar haciendo que que el consumo de tiempo haga prohibitivo su uso.

V-A.1. k-Means: Esta es una técnica de agrupamiento, la cual asume que los datos (del tipo que sean) está representados como vectores en algún espacio $n-$dimensional y con esto pueden utilizar una medida de similitud basada en el ángulo que puedan formar dos vectores en dicho espacio.

La técnica de k-means lo que hace inicialmente es seleccionar k documentos "semilla" arbitrarios. Luego, cada documento de entrada es asignado al documento semilla más similar. Luego, los documentos semilla son recalculados para ser el centroide de todos los documentos asignados a dicha semilla. Este proceso es realizado hasta que las coordenadas de la semilla se estabilicen. El principal problema de esta técnica es que cuando la dimensionalidad es muy alta, los tiempos de respuesta crecen, y este es en general el caso de textos.

V-A.2. Agrupamiento aglomerativo: En esta técnica los documentos son fusionados en super-documentos o grupos[24] hasta que solo quede un grupo. La idea de esta técnica es la de generar jerarquías basadas en dichas fusiones, de modo que se pueden establecer similaridades

entre documentos.

V-B. Técnicas probabilísticas

Estas técnicas se basan en la siguiente supocisión: se tienen m clases de tópicos, $c_1, c_2, ..., c_3$ con algunos documentos semilla D_c asociados a cada una donde, la probabilidad de que un documento pertenezca a una clase está dado por : $\frac{|D_c|}{\sum_{\forall}|D_c|}$. Con esta primera aproximación a la probabilidad de documentos se pueden comenzar a explicar algunas técnicas probabilísticas.

V-B.1. 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:

$$\Pr(c_i \mid s_j) = \frac{\Pr(c_i)\Pr(s_j \mid c_i)}{\Pr(s_j)}$$

donde s_j es la muestra j y $c_i$es la clase i.

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 [24]. Se asume que para cada clase c existe un modelo generador de textos. Dicho modelo tiene como parámetro el valor $\phi_{c,t}$, el cual indica la probabilidad de que un documento en la clase c contenga el término t al menos una vez. De esta forma, se define la probabilidad de que un documento d pertenezca a una clase c, $\Pr(d \mid c)$ como:

$$\arg\max\left(\Pr(d \mid c) = \prod_{t\in d}\phi_{c,t}\prod_{t\in T, t\notin d}(1-\phi_{c,t})\right)$$

donde T es el universo o vocabulario de términos.

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.

V-B.2. 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 [25]:

- 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.
- 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 [25].

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 en [26].

V-C. Técnicas bio-inspiradas

Estos métodos se encuentran en el campo de la computación bio-inspirada, cuya idea es utilizar analogías con sistemas naturales o sociales para diseñar métodos heurísticos no determinísticos de búsqueda, aprendizaje y comportamiento entre otros. Dentro de estos métodos se pueden identificar algunas aproximaciones, cada una de las cuales, buscan emular algún sistema natural. A continuación se explican tres de ellos: redes neuronales, sistemas inmunológicos artificiales y los método evolutivos. Los primeros son divididos en dos ya que cada uno de ellos tiene métodos de entrenamiento diferentes.

V-C.1. 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: [27] y [28]. Adicionalmente, se pueden encontrar algunos algoritmos de entrenamiento supervisado y no supervisado en [29]

V-C.2. 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 [30]. 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 [31] 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 [32], 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 [31], 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 [30] y [31].

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 (Sección V-A) y luego aplicar el método específico.

V-C.3. 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[33] o la selección clonal [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 solo poder clasificar en dos clases podría 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, en [35], se 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, en [36] se 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.

V-C.4. 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 [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.

V-D. 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 [38], 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.

VI. Resumen

En este trabajo se ha planteado el problema de relacionar el conocimiento tácito extraido a partir del análisis de redes sociales creadas utilizando la información de correos electrónicos y el manejo del conocimiento explícito, el cual es obtenido a partir de la información extraida de los documentos contenidos en repositorios personales así como en los cuerpos de mensajes de correo electrónico.

Luego de plantear el problema, se explica en que consiste la gestión de conocimiento y los tipos de conocimiento con los que se puede trabajar, así como las formas en que se realiza su gestión actualmente. Es claro, que en el caso del conocimiento tácito, la mejor forma de gestionarlo, es utilizando el análisis de redes sociales, ya que con estas se puede descubrir el flujo de la información así como una

medida de cuan dispersa está. En el caso del concimiento explícito, existen diversas formas de extraerlo, las cuales pueden ser vistas como agrupación o categorización y clasificación, el problema de esto, es que la técnica a utilizar está muy ligada al problema particular que se desea resolver, por lo que puede ser importante tener en cuenta la posibilidad de combinar algunos métodos con el fin de obtener los mejores resultados al enfrentarse a problemas genéricos.

Dentro de las soluciones revisadas, solo aparecen aquellas que gestionan conocimiento tácito o explícito, pero no ambos, es decir, no hay una que los relacione

VII. Preguntas de Investigación

A continuación plantearé tres preguntas de investigación que surgen luego del trabajo previo.

VII-A. ¿Cómo relacionar el conocimiento explícito con el conocimiento tácito incluido en una red social?

Actualmente las empresas y las personas, poseen una gran cantidad de documentos en variados formatos, que pueden buscar y descargar de Internet, crear o recibir a través de un medio como el correo electrónico. Estos repositorios tienen una gran cantidad de información, la cual es totalmente inútil si se deja allí, de hecho, solo sirven para ocupar espacio en los medios de almacenamiento. En cambio, si se realiza una buena gestión, se podrá obtener conocimiento de ellos, que para este caso, puede ser información sobre la información, lo que de forma inmediata le va a dar un valor agregado para las personas que los posean.

Adicionalmente, las personas poseen un conocimiento que reposa en el flujo de información que tienen con otras personas, las cuales son generalmente cercanas o que se interesan en algún tema común. Este conocimiento le puede decir a las personas sobre que temas hablan o sobre que temas manejan un cierto nivel de experticia.

Hasta ahora estos dos tipos de conocimiento han sido utilizados por separado, lo que plantea la posibilidad de buscar una integración entre ellos, de modo que se le pueda entregar el mayor potencia a un usuario y reducir el conocimiento que se pierde día a día.

VII-B. ¿Cuál método de extracción de conocimiento explícito es el mejor para manejo de documentos personales?

Los documentos que posee una persona, ya sean descargados, guardados o creados, pueden ser sobre diversos temas, los cuales pueden o no, ser totalmente conocidos. Pero, suponiendo que todos los temas son conocidos, es muy probable que diferentes documentos guarden algún tipo de relación temática y porbablemente el usuario no lo sepa.

Es por esto que toma importancia el hecho de elegir correctamente el método de extracción de conocimiento de los documentos, de modo tal que se pueda hacer una categorización para encontrar los temas existentes y luego una clasificación de documentos nuevos, permitiendo así crear un mapa conceptual documental.

VII-C. ¿Es posible realizar recomendaciones de colaboración en una red social personal?

Una de las aplicaciones más difundidas del análisis de redes sociales, son los sistemas de recomendación. Estos sistemas están diseñados para realizar recomendaciones a un usuario sobre personas dentro de una red social corporativa que lo pueden desde ayudar a resolver un determinado problema hasta el diseño de equipos de trabajo para nuevos proyectos. Esto es posible en gran medida por el tamaño que puede tener una red social en una empresa, así como la diversidad de personas que en ella trabajan. Pero, cuando la red social es personal, por lo que en principio es reducida, ¿como se puede utilizar para recomendar a algún miembro de dicha red para efectuar algún tipo de colaboración?, es probable que haya la necesidad de tener en cuenta otros factores que no se tienen en cuenta en la empresas, es decir, teniendo en cuenta no solo mi red social, sino también la información del conocimiento documental se podrían realizar recomendaciones en una red de este tamaño.

References

[1] I. Spiegler, "Knowledge management: a new idea or a recycled concept?," *Commun. AIS*, vol. 3, no. 4es, p. 2, 2000.

[2] 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.

[3] 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.

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

[5] W. F. Cody, J. T. Kreulen, V. Krishna, and W. S. Spangler, "The integration of business intelligence and knowledge management," *IBM Syst. J.*, vol. 41, no. 4, pp. 697–713, 2002.

[6] V. Marik, M. Pechouek, and O. Stepankova, "Social knowledge in multi-agent systems," in *Multi-Agent Systems and Applications : 9th ECCAI Advanced Course ACAI 2001 and Agent Link's 3rd European Agent Systems Summer School, EASSS 2001*, (New York, NY, USA), pp. 211–245, Springer-Verlag New York, Inc., 2001.

[7] G. Schreiber, H. Akkermans, A. Anjewierden, R. de Hoog, N. Shadbolt, W. V. D. Velde, and B. Wielinga, *Knowlede Engineering and Management: The CommonKADS Methodology.* MIT Press, 1999.

[8] C. S. Campbell, P. P. Maglio, A. Cozzi, and B. Dom, "Expertise identification using email communications," in *CIKM '03: Proceedings of the twelfth international conference on Information and knowledge management*, (New York, NY, USA), pp. 528–531, ACM Press, 2003.

[9] J. Goodman and V. R. Carvalho, "Implicit queries for email," in *Proceedings of Second Conference on Email and Anti-Spam CEAS 2005*, (Stanford University), July 21 & 22 2005.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

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

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

[26] 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.

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

[28] 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.

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

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

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

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

[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] 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.