Generación, análisis y utilización de redes sociales en correos electrónicos personales

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

***Abstract*— 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 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.**

I. Introducción

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. 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 explican algunas técnicas relevantes para la extracción de conocimiento y la clasificación de documentos, en la sección III, se explican algunos tópicos generales sobre gestión de conocimiento, en la sección IV se explican generalidades sobre las redes sociales, las formas en que se pueden generar y las medidas para su análisis, en la sección V, se explican dos aplicaciones derivadas del análisis de redes sociales, finalmente, se hace un resumen de lo explicado en el trabajo.

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

II. 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.

II-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 [1] 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* [1] (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 [1] 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 [1] (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.

II-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.

II-A.2. Agrupamiento aglomerativo: En esta técnica los documentos son fusionados en super-documentos o grupos[1] 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.

II-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 : $\frac{|D_c|}{\sum_c |D_c|}$. Con esta primera aproximación a la probabilidad de documentos se pueden comenzar a explicar algunas técnicas probabilísticas.

II-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:

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

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 [1]. 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, $\Pr\left(d \mid c\right)$ como:

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

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.

II-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 [2]:

- 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 [2].

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 [3].

II-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.

II-C.1. 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 es supervisado, por lo que para cada patrón de entrada se necesita 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: [4] y [5]. //TODO

II-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 [6]. 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 [7] 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 [8], 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 [7], 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 [6] y [7].//TODO

II-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[9] o la selección clonal [10].

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 [11], 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 [11]. Una vez entrenados los detectores, el sistema está en capacidad de clasificar cualquier documento nuevo que se le presente.

II-C.4. Computación evolutiva: La computación evolutiva recoge varios métodos de optimización que basan su funcionamiento en la teoría de 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.

Trabajos interesantes: [12], [11]....//TODO

III. Gestión de Conocimiento

La gestión del conocimiento tiene variadas definiciones, depende del punto de vista desde donde se den. En [13] 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 [13].

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 [14] 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 [15] 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 //TODO:

III-A.2. No articulable //TODO:

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.

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.

V. Aplicaciones

Las redes sociales y su análisis tienen varias aplicaciones. Las primeras se refieren a la gestión del capital intelectual en forma de conocimiento tácito. Otra es la de mejorar la comunicación entre los empleados de una compañía de modo que se puedan establecer redes de colaboración y así poder 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.

A continuación se explican dos de las aplicaciones del análisis de las redes sociales: trabajo colaborativo y gestión de correos.

V-A. 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 [16].

V-B. 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 [17] 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 [18], 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.

Otras técnicas propuestas en: [19], [20], [21], [22]

Todas las técnicas anteriores utilizan técnicas basadas en redes sociales, sin embargo, existen iniciativas diferentes, como la propuesta por [23], en la que se plantea la utilización de un modelo basado en sistemas inmunológicos artificiales.

(Falta explicar los mecanismos)

VI. Resumen

En este trabajo se han explicado algunas técnicas para realizar clasificación y extracción de conocimiento de documentos y textos en general, lo cual es una tarea básica a la hora de gestionar el conocimiento explícito de una organización o de una persona, saber que información documental posee y como se puede organizar y/o clasificar de modo que se puedan relacionar entre ellos. Luego se explicaron algunas generalidades de la gestión de conocimiento, lo cual da una base para saber que es lo que se busca al realizar tal gestión, en este caso, se explicó el conocimiento tácito y el conocimiento explícito. Para el primero aún no es claro como se debe adquirir y gestionar, para el segundo, se explicaron algunas técnicas en la sección II, pero lo que no se ha explicado es como conectarlos. Luego se explicó la teoría sobre redes sociales, como generarlas y algunas de las medidas que se tienen para su análisis.

Es importante notar que si al análisis de las redes sociales, se les añade la información sobre documentos, se puede tener un modelo de información más completo que va a permitir responder más preguntas acerca del comportamiento organizacional.

References

[1] S. Chakrabarti, "Data mining for hypertext: a tutorial survey," *SIGKDD Explor. Newsl.*, vol. 1, no. 2, pp. 1–11, 2000. Este es un "survey"sobre diferentes métodos para extraer conocimiento de textos, específicamente de hipertextos. El autor divide la discusión en cinco partes: * Modelos básicos * Aprendizaje supervisado * Aprendizaje no supervisado * Aprendizaje Semi-supervisado * Análisis de redes sociales Este artículo es bueno, es un poco corto para ser un survey, sin embargo muestra las técnicas más utilizadas para realizar búsquedas de y entre documentos, es importante. No tiene en cuanta cosas como la maldición de la dimensionalidad.

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

[3] 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.

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

[5] 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.

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

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

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

[9] 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.

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

[11] 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.

[12] 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.

[13] I. Spiegler, "Knowledge management: a new idea or a recycled concept?," *Commun. AIS*, vol. 3, no. 4es, p. 2, 2000. Este no es un artículo, es más que eso. Es un estudio sobre la gestión del conocimiento, cuales spn sus fases, que estudios se han hecho sobre eso. Tiene diferentes puntos de vista: datos e información, capital intelectual y mapas de conocimiento, tecnología y su influencia sobre el conocimiento, modelos y el proceso de transformación del conocimiento. Es interesante, ya que no es un enfoque desde el punto de vista de ciencias de la computación, sino desde la administración y las ciencias de la información. Es definitivamente un documento clave para armar la parte sobre gestión del conocimiento del marco teórico.

[14] 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. En este artículo crean redes sociales de conocimiento pero utilizando el flujo de conocimiento tácito. Esto no es una tarea simple; para capturar dicho conocimiento diseñaron unos formularios con ciertas preguntas muy específicas. Personalmente creo que formularios no es la forma de extraer conocimiento tácito, sin embargo, el trabajo es muy interesante ya que hace una división muy clara entre los tipos de conocimieto y además añade al conocimiento explícito una división: el conocimiento codificado, que es todo aquel que está escrito de alguna forma. Más adelante utiliza el análisis de redes sociales para analizar las redes de conocimiento tácito que han generado, tienen resultados interesantes.

[15] 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. Este artículo cuenta de forma clara que es el conocimiento tácito o suave, y realiza un gran énfasis en el hecho de que la gestión del conocimiento que se hace en el mundo occidental es muy diferentes al que se realiza en el mundo oriental. Hay una bibliografía extensa, en domde nombra autores muy reconocidos en materia de gestión de conocimiento, como Nonaka y Takeuchi. Algo muy interesante que hacen es crear una tabla con todo aquello que puede ser tratado como conocimiento tácito, pero además, la dividen en el conocimiento

tácito que puede ser articulado y el que no. Luego proponen como herramienta para gestionar el conocimiento, el análisis de redes sociales, de donde obtienen información sobre el flujo de conocimiento en una organización y como puede estar este entorno a una persona o portero, como lo llaman ellos. Es muy interesante, creo que realiza aportes muy interesantes en la teoría sobre gestión de conocimiento y muestra un caso de estudio general para el análisis de redes sociales.

[16] 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. Este artículo está enfocado al trabajo colaborativo asistido por computador (CSCW por sus iniciales en inglés). Y muetra como recomendar colaboraciones utilizando las redes sociales y su análisis. Es un artículo que no tienen mucho fundamento matemático (no quiere decir que sea malo) ya que está visto desde el concepto de organización y los grupos que se pueden forma allí y no desde el análisis matemáticos de grafos. Es teóricamente fuerte, por lo que puede ayudar en la construcción del estado del arte tanto para larte de redes sociales como para la parte de trabajo colaborativo.

[17] 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.

[18] 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. Este artículo muestra otra forma para realizar la clasificicación de correos electrónico, pero basada en la reputación de las personas de la red. Esta reputación es econtrada a partir de las relaciones existentes en la red, no solo las de un grado, sino más, lo que permite establecer si una persona con la que no se está conectada directamente es o no confiable. Para demostrarlo, desarrollaron un prototipo cliente de correo, el cual está en capacidad de encontrar la confiabilidad de un mensaje. Es muy interesante, no solo provee información para el marco teórico sino que aporta información sobre la aplicaciones que tiene.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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. Este artículo presenta una forma de clasificación automática de correos utilizan sistemas inmunológicos artificiales. Es muy interesante. Lo que se propone es un esquema de selección clonal, en donde se utiliza el esquema de entrenamiento utilizado comunmente. Los correos nuevos son tratados como antígenos, que luego son presentados a todo el sistema inmunológico. Algo muy interesante es la representación de las células, en donde solo se tienen en cuenta las palabras que van en el remitente y en el asunto, logrando que el número de células no sea muy alto (algo interesante para un AIS). Hay que estudiarlo con más profundidad y revisar otros por el estilo.