



UNIVERSIDAD NACIONAL DE COLOMBIA
SEDE BOGOTÁ

FACULTAD DE INGENIERÍA
UNIDAD DE POSGRADOS

PRESENTACIÓN PROPUESTA

(Utilice o amplíe los espacios sugeridos según convenga – Tamaño Fuente: 12, Interlineado: 1,5)

TESIS DE MAESTRÍA: ☒ TRABAJO FINAL DE ESPECIALIZACIÓN: ☐

1. PROPONENTE: Juan David Cruz Gómez _____

CÓDIGO: 299618_____

2. PROGRAMA: Maestría en Ingeniería de Sistemas y Computación

3. DIRECTOR PROPUESTO: Ing. Fabio González Osorio, PhD _____

DEPARTAMENTO: Ingeniería de Sistemas e Industrial _____

ASESORES: _____

4. TÍTULO: Sistema para identificar, extraer y visualizar conocimiento en
correos electrónicos

ÁREA: Gestión del Conocimiento-_____

LÍNEA DE INVESTIGACIÓN: _____

5. ANTECEDENTES Y JUSTIFICACIÓN: La visión del conocimiento como un activo corporativo ha despertado el interés de las grandes compañías por buscar y desarrollar herramientas que le permitan gestionar dicho activo de la mejor forma posible. Para esto, se ha desarrollado la gestión de conocimiento y dentro de esta, diversos conceptos, como los diferentes tipos de conocimiento y herramientas para su manejo adecuado, como el análisis de redes sociales. Sin embargo, estos conceptos han salido de los límites corporativos para convertirse en poderosas herramientas que

facilitan la búsqueda de diferentes tipos de información así como el trabajo colaborativo. De esta forma, se han desarrollado técnicas basadas en teorías de gestión de conocimiento 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.

Una de las herramientas más poderosas para encontrar y estudiar las relaciones entre organizaciones y/o personas, es el análisis de redes sociales. 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 el análisis de redes sociales ya es una herramienta bastante útil, la información que aporta puede ser potenciada si se le añade la información sobre los documentos y sus relaciones que un individuo posee sobre diferentes temas.

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, 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 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.

- a. Conocimiento tácito: El conocimiento tácito o implícito es aquel que poseen las personas y que no puede ser fácilmente extraído. 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 aprender y se va a caer; solo la experiencia le permitirá interiorizar dicho conocimiento. 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. 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. El primero se refiere a 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. El segundo, se refiere a 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.
- b. Conocimiento explícito: El conocimiento explícito es todo aquel conocimiento que está disponible en documentos, bases de datos o cualquier otro tipo de repositorio físico, que además, describe algún procedimiento o técnica. Esto quiere decir que puede ser aprendido

a través de libros o de artículos y es fácilmente extraíble y administrable.

- c. Ciclo de creación del conocimiento: Según Nonaka [4], el conocimiento dentro de una organización se genera a partir de cuatro fases:
- i. 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.
 - ii. 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.
 - iii. 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.
 - iv. 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. [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 con posterioridad. Por otro lado, la difusión, consiste en fomentar la comunicación entre individuos que componen la organización de modo que se pueda crear el conocimiento tácito colectivo.

- d. Extracción de Conocimiento en Correos Electrónicos: 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, las que permiten generar más

conocimiento cada instante.

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 muy rápidamente. 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 mundo¹.

Lo anterior hace que actualmente las empresas y las personas posean una gran cantidad de información contenida en mensajes de correo electrónico, los cuales pueden o no, tener archivos adjuntos relacionados con un tema específico. Adicionalmente, estos mensajes contienen información que no es evidente, pero, que si se extrae y presenta de la manera correcta, le puede aportar conocimiento al usuario final que la requiera.

Para realizar la extracción, existen variadas técnicas que pueden ser clasificadas según el método que utilizan para realizar la categorización del texto. Pueden ser determinísticas, que buscan "vectorizar" el texto a ser estudiado: el método utilizado 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. En [5] y en [6] se pueden ver estudios sobre este tipo de técnicas. Otro tipo son las probabilísticas, en donde se calcula la probabilidad de que un documento pertenezca a determinada clase. Estas técnicas se explican en [5], [7], [8] y en [9]. Las siguientes técnicas son las llamadas bio-inspiradas, estas utilizan 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. Algunos de ellos son: las redes neuronales, sistemas inmunológicos artificiales y los métodos evolutivos. Dos

¹ Fuente: <http://archives.cnn.com/TECH/>.

trabajos interesantes pueden ser encontrados en [10], en el que plantea la utilización de un algoritmo genético paralelo para extraer información que le permita realizar tareas de vigilancia estratégica sobre Internet, y en [11], donde 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. Finalmente, existen otras técnicas, que no caben en la división realizada, por ejemplo la propuesta en [12], 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.

En [13] se propone una forma de extraer información de documentos que están relacionados entre si a través de un hilo conductor. Un tipo especial de estos documentos son los correos. La idea se basa en la técnica de extracción de información que busca las características recurrentes en un texto.

Actualmente se está difundiendo una técnica de análisis conceptual de documentos llamada Análisis Formal de Conceptos (FCA, por sus iniciales en inglés) desarrollado por Wille en 1982 [14], que busca modelar de forma matemática los conceptos contenidos en textos y presentar la relación de sus conceptos en forma de un grafo, le cual es generado a partir de ciertos atributos; una explicación muy interesante de los conceptos involucrados puede encontrarse en [14, 15, 16].

En [17] se propone un sistema de clasificación de colecciones de correos electrónicos basado en FCA. En este trabajo se expone la dificultad del trabajo con correos electrónicos debido a la informalidad del lenguaje utilizado y la alta contextualización del mismo; sin embargo, algo que merece más atención que el clasificador en si mismo, es la forma en que se debe analizar el grafo para extraer la información correcta.

6. IDENTIFICACIÓN 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 entre personas y que permitan el crecimiento intelectual de las mismas.

Este es el caso de la explotación de los dos tipos de conocimiento: tácito y explícito. El primero, puede mostrar información acerca del flujo de información acerca de ciertos temas de uso predilecto del usuario, así como, información no evidente en el mensaje, por ejemplo, información sobre un archivo adjunto (no análisis del documento adjunto) pero que puede ser útil. El segundo, mostrará información acerca del contenido del mensaje, realizando algún análisis de su contenido desde el punto de vista semántico de modo que se puedan extraer conceptos útiles para relacionarlos con otros similares. Esta información extraída, es más útil en la medida en que esté relacionada, es decir, que los dos tipos de conocimiento estén relacionados de alguna manera, de modo tal que cada uno ellos apoye al otro en la cantidad de información que puede aportar, ya sea para que el usuario la utilice directamente, por ejemplo, para buscar expertos en un tema o para que otra herramienta le de un valor agregado, por ejemplo, para realizar algún tipo de recomendación.

El conocimiento explícito, representado en documentos y textos, también contiene un conocimiento tácito, el cual ha sido añadido a lo largo de su creación por uno o más autores. Esto se puede ver más claro en los documentos generados durante la elaboración de un proyecto, en donde diferentes personas aportan a su creación, siendo este un objeto clave en la interacción de dichas personas.

Para realizar esto, se debe buscar un modelo de gestión de conocimiento que permita trabajar con información personal y tal vez, con información que esté a un grado de distancia dentro de la red social, lo que lo hace interesante, ya que la mayoría de modelos trabajan con redes bastante estables y que cambian muy poco en el tiempo, como la red de un tema posicionado en una compañía. Dichas redes representan entre otras cosas, el flujo de información entre dos o más individuos u organizaciones y su análisis ayuda a la identificación de ciertos puntos en esta que pueden ser vulnerables o de gran ayuda al momento de planear y/o ejecutar un proyecto determinado. Estas redes se construyen a partir de la extracción de la información de origen y destino contenida en los mensajes de correo

electrónico y del conocimiento contenido en documentos que no está evidentemente relacionado con los individuos.

7. OBJETIVO GENERAL Y OBJETIVOS ESPECÍFICOS:

a. Objetivo General: Desarrollar un sistema que permita extraer conocimiento útil de correos electrónicos y documentos asociados con el fin de gestionarlo de forma que represente un valor adicional para el usuario.

b. Objetivos Específicos

i. Extraer la información relevante de los contactos y el conocimiento de los mensajes de correo electrónico así como de los archivos adjuntos.

ii. Desarrollar un modelo de gestión de conocimiento que permita asociar objetos intermediadores a las relaciones de las redes sociales encontradas.

iii. Desarrollar un prototipo de software que permita extraer, visualizar y analizar la información de los correos y las redes contenidas en ellos.

8. METODOLOGÍA: La primera fase del trabajo consiste en realizar una rápida revisión bibliográfica de las técnicas de extracción de conocimiento en textos, específicamente en correos electrónicos; esto además de incluir el cuerpo de los mensajes también incluye los datos de origen y destino, lo que permite realizar otros estudios sobre los mensajes. Adicionalmente se hará una revisión de algunas técnicas de gestión de conocimiento explícito y tácito, incluyendo la generación y análisis de redes sociales. Todo lo anterior es enfocado a actualizar el estado del arte que se ha escrito.

Luego de la actualización del estado del arte, se debe plantear un modelo de gestión de conocimiento que permita asociar diferentes objetos intermediadores con las redes sociales identificadas. Para esto se deben identificar los posibles objetos que están presentes en un contexto personal, específico en el cuál los mensajes son generados. Este modelo debe ser validado por un experto en el área y por un experto en el contexto, que en este caso, será el dueño de los mensajes de correo electrónico.

Una vez creado el modelo, se comienza el desarrollo de una

representación estructurada de los mensajes de correo electrónico, de modo que sea sencillo manipularlos, es decir, almacenarlos y extraer la información que ellos contienen de manera eficiente. El conjunto de mensajes de correo utilizado es el de ENRON², el cuál fue hecho público durante la investigación que realizó la Comisión Federal de Regulación de Energía de los Estados Unidos. Este conjunto es interesante porque es de la vida real, es decir, fue generado por los empleados de la compañía en un espacio de tiempo determinado; además, ha sido utilizado para comprobar el funcionamiento de modelos de categorización de correos en diferentes conferencias internacionales, por lo que sirve como referente de comparación de los resultados que este trabajo arroja.

Una vez se tienen los mensajes almacenados en un formato estructurado, se puede extraer la información relevante de origen, destino y demás meta-datos, como, asunto, fecha, cuerpo y archivos adjuntos para realizar los análisis respectivos: con la información de origen y destino, se construyen las matrices de adyacencia que son utilizadas posteriormente para generar el grafo que representa la red social. La información del cuerpo y de los archivos adjuntos, se categoriza de modo que sirvan como objeto intermediador en las relaciones entre personas.

Las redes sociales generadas, corresponden a las redes en las que una persona es el principal origen y principal destino, es decir, es el dueño de los mensajes que le llegan y de los que envía. Esto sugiere que esta persona debe estar en el centro de la red y esta se debe expandir alrededor de dicha persona. Es necesario estudiar algunas opciones de representación gráfica que permita ver la red de esta forma, asignando además algún tipo de importancia al contacto y la conexión.

Finalmente, se generará un prototipo que integra los desarrollos realizados durante las fases iniciales del proyecto, siguiendo una metodología orientada a prototipos. Se generarán tres prototipos, el primero, será el encargado de estructurar los mensajes de correo y de extraer su información, el segundo, será el encargado de crear, analizar y representar las redes sociales; estos prototipos se construyen aparte, de modo que sus procesos de desarrollo no interfieren. El tercer prototipo será la integración preliminar de los dos anteriores y sobre el cuál se realizarán las pruebas de funcionamiento generales.

² <http://www.isi.edu/~adibi/Enron/Enron.htm>

Las pruebas del software serán llevadas a cabo de dos formas: pruebas de unidad, que buscan comprobar errores lógicos y de programación en los prototipos, y pruebas de funcionamiento, en las que un experto validará los resultados obtenidos en la integración de objetos y relaciones.

9. ACTIVIDADES A DESARROLLAR:

- a. Revisión y actualización del estado del arte
 - i. Métodos de extracción de conocimiento de correos electrónicos
 - ii. Gestión de Conocimiento tácito y explícito
 - iii. Análisis de redes sociales
- b. Desarrollo del modelo de gestión de conocimiento
 - i. Identificación de objetos intermediadores
 - ii. Identificación de las relaciones entre objetos intermediadores y redes sociales.
 - iii. Validación del modelo
- c. Desarrollo de Prototipos
 - i. Desarrollo del prototipo I
 - 1. Especificación (Entradas/Salidas)
 - 2. Arquitectura de software
 - 3. Análisis de requerimientos
 - 4. Diseño
 - 5. Implementación
 - a. Revisión de librerías utilitarias
 - b. Estructuración de mensajes
 - c. Extracción de datos
 - d. Extracción de conocimiento
 - 6. Pruebas
 - a. Funcionamiento global
 - b. Unidad
 - ii. Desarrollo del prototipo II
 - 1. Especificación (Entradas/Salidas)
 - 2. Arquitectura de software
 - 3. Análisis de requerimientos
 - 4. Diseño
 - 5. Implementación

11. BIBLIOGRAFÍA BÁSICA:

- [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. Takeuchi, *The knowledge creating company*. Oxford University Press, 1995.
- [5] S. Chakrabarti, "Data mining for hypertext: a tutorial survey," *SIGKDD Explor. Newsl.*, vol. 1, no. 2, pp. 1–11, 2000.
- [6] J. A. Black and N. Ranjan, "Automated event extraction from email." Final Report of CS224N/Ling237 Course in Stanford: <http://nlp.stanford.edu/courses/cs224n/2004/>, Spring 2004.
- [7] M. Sahami, S. Dumais, D. Heckerman, and E. Horvitz, "A bayesian approach to filtering junk Email," in *Learning for Text Categorization: Papers from the 1998 Workshop*, (Madison, Wisconsin), AAAI Technical Report WS-98-05, 1998.
- [8] I. L. MacDonald and W. Zucchini, *Hidden Markov and Other Models for Discrete-Valued Time Series*. Chapman & Hall/CRC, 2002.
- [9] S. Sarawagi and W. W. Cohen, "Semi-markov conditional random fields for information extraction," in *Advances in Neural Information Processing Systems*, December 13 2004.
- [10] 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.
- [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] 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.
- [13] D. Masterson and N. Kushmerick, "Information extraction from multi-document threads," in Proceedings of the International Workshop on Adaptive Text Extraction and Mining, (Catvat – Dubrovnik (Croatia)), September 22 2003.
- [14] U. Priss, "Formal concept analysis in information science," Annual Review of Information Science and Technology, ARIST, vol. 40, pp. 521 – 543, 2006 (To be published).
- [15] U. Priss, "A graphical interface for document retrieval based on formal concept analysis," in Proceedings of the 8th Midwest Artificial Intelligence and Cognitive Science Conference (E. Santos, ed.), (Dayton Ohio), pp. 66 – 70, Wright State University, May 30 – June 1 1997.
- [16] P. Cimiano, A. Hotho, G. Stumme, and J. Tane, "Conceptual knowledge processing with formal concept analysis and ontologies," in Proceedings of the The Second International Conference on Formal Concept Analysis (ICFCA 04) (Springer, ed.), pp. 189 – 207, 2004.
- [17] I. Richard J. Cole and P. W. Eklund, "Analyzing an email collection using formal concept analysis," in PKDD '99: Proceedings of the Third European Conference on Principles of Data Mining and Knowledge Discovery, (London, UK), pp. 309–315, Springer-Verlag, 1999.

12. RECURSOS FÍSICOS:

- a. Equipos de cómputo (Proponente)
- b. Acceso a Internet
- c. Suministros de papelería
- d. Bases de datos bibliográficas y repositorios de datos de prueba

13. COSTOS DEL TRABAJO Y FUENTES DE FINANCIACIÓN:

Concepto	Costo estimado	Fuente de Financiación	
		Universidad Nacional	Recursos Propios
Costo de personal			
Estudiante	\$ 12.000.000		\$ 12.000.000
Director de trabajo de grado	\$ 16.000.000	\$ 16.000.000	
Costos operativos			
Equipo de cómputo	\$ 3.500.000		\$ 3.500.000
Materiales y Suministros (Papelería)	\$ 400.000		\$ 400.000
Material bibliográfico impreso	\$ 300.000		\$ 300.000
Material bibliográfico digital	\$ 150.000		\$ 150.000
Realización de publicaciones	\$ 2.500.000	\$ 1.000.000	\$ 1.500.000
	Sub-Total	\$ 17.000.000	\$ 17.850.000
	TOTAL		\$ 34.850.000

14. COMENTARIO CON VISTO BUENO DEL DIRECTOR: (calificar los siguientes aspectos: organización, pertinencia, relevancia y originalidad).

15. FIRMA DEL PROPONENTE

16. FIRMA DEL DIRECTOR (ASESORES)

17. FECHA