

Gestión del Conocimiento Contenido en Correos Electrónicos

Extracción y representación de información no evidente

Ing. Juan David Cruz Gómez, 299618
Propuesta de Trabajo de Grado

Seminario de Investigación II

Agenda

- 1 Antecedentes y Justificación
- 2 Planteamiento del Problema
- 3 Objetivos
 - Objetivo General
 - Objetivos Específicos
- 4 Metodología
- 5 Actividades a Desarrollar
- 6 Bibliografía

Antecedentes y Justificación (1)

- En la actualidad, el conocimiento es visto como un activo mas.
- Debe ser gestionado de la mejor forma posible con el fin de agregar valor a los productos y servicios prestados por una compañía.
- Se han creado diferentes técnicas
 - Extracción de conocimiento explícito
 - Modelos de identificación y codificación de conocimiento tácito
 - Análisis de redes sociales
- La mayoría de estudios se han creado a partir de casos de aplicación en empresas de consultoría

Antecedentes y Justificación (1)

- En la actualidad, el conocimiento es visto como un activo mas.
- Debe ser gestionado de la mejor forma posible con el fin de agregar valor a los productos y servicios prestados por una compañía.
- Se han creado diferentes técnicas
 - Extracción de conocimiento explícito
 - Modelos de identificación y codificación de conocimiento tácito
 - Análisis de redes sociales
- La mayoría de estudios se han creado a partir de casos de aplicación en empresas de consultoría

Antecedentes y Justificación (1)

- En la actualidad, el conocimiento es visto como un activo mas.
- Debe ser gestionado de la mejor forma posible con el fin de agregar valor a los productos y servicios prestados por una compañía.
- Se han creado diferentes técnicas
 - Extracción de conocimiento explícito
 - Modelos de identificación y codificación de conocimiento tácito
 - Análisis de redes sociales
- La mayoría de estudios se han creado a partir de casos de aplicación en empresas de consultoría

Antecedentes y Justificación (1)

- En la actualidad, el conocimiento es visto como un activo mas.
- Debe ser gestionado de la mejor forma posible con el fin de agregar valor a los productos y servicios prestados por una compañía.
- Se han creado diferentes técnicas
 - Extracción de conocimiento explícito
 - Modelos de identificación y codificación de conocimiento tácito
 - Análisis de redes sociales
- La mayoría de estudios se han creado a partir de casos de aplicación en empresas de consultoría

Antecedentes y Justificación (2)

- Existen varios estudios y propuestas sobre como se debe codificar el conocimiento tácito
 - ¿Es realmente codificable?
 - ¿Cual lo es y cual no?
 - ¿Cómo puede ser transmitido?
- La mayoría de estudios en conocimiento explícito se enfoca a la gestión documental
 - Tecnicas de categorización
 - Identificación de temas clave
 - Bases documentales indexadas
- El correo electrónico se ha convertido en uno de los medios de comunicación más importantes

Antecedentes y Justificación (2)

- Existen varios estudios y propuestas sobre como se debe codificar el conocimiento tácito
 - ¿Es realmente codificable?
 - ¿Cual lo es y cual no?
 - ¿Cómo puede ser transmitido?
- La mayoría de estudios en conocimiento explícito se enfoca a la gestión documental
 - Tecnicas de categorización
 - Identificación de temas clave
 - Bases documentales indexadas
- El correo electrónico se ha convertido en uno de los medios de comunicación más importantes

Antecedentes y Justificación (2)

- Existen varios estudios y propuestas sobre como se debe codificar el conocimiento tácito
 - ¿Es realmente codificable?
 - ¿Cual lo es y cual no?
 - ¿Cómo puede ser transmitido?
- La mayoría de estudios en conocimiento explícito se enfoca a la gestión documental
 - Tecnicas de categorización
 - Identificación de temas clave
 - Bases documentales indexadas
- El correo electrónico se ha convertido en uno de los medios de comunicación más importantes

Antecedentes y Justificación (3)

- Muestra vínculos de comunicación entre diferentes personas u organizaciones
- Contiene información no evidente para los usuarios
 - Vínculos con otros individuos
 - Hilos de conversación
 - Objetos intermediadores

Antecedentes y Justificación (3)

- Muestra vínculos de comunicación entre diferentes personas u organizaciones
- Contiene información no evidente para los usuarios
 - Vínculos con otros individuos
 - Hilos de conversación
 - Objetos intermediadores

Planteamiento del Problema.

- El conocimiento explícito, representado en forma de documentos y textos contiene una porción de conocimiento tácito, añadido a través de su procesamiento e creación y modificación.
- Esto representa un objeto intermediador entre todas las personas que han intervenido.
- Se debe tomar cada mensaje de correo electrónico para obtener la información de contactos para generar las redes sociales.
- Luego, se deben analizar los datos relevantes que están en el encabezado así como el o los archivos adjuntos.
- Finalmente crear una relación entre esos tres factores para entregar conocimiento adicional al usuario.

Outline

- 1 Antecedentes y Justificación
- 2 Planteamiento del Problema
- 3 Objetivos**
 - Objetivo General
 - Objetivos Específicos
- 4 Metodología
- 5 Actividades a Desarrollar
- 6 Bibliografía

Objetivo General

Extraer conocimiento útil de correos electrónicos y documentos asociados con el fin de gestionarlo de forma que represente un valor adicional para el usuario.

Outline

- 1 Antecedentes y Justificación
- 2 Planteamiento del Problema
- 3 Objetivos**
 - Objetivo General
 - **Objetivos Específicos**
- 4 Metodología
- 5 Actividades a Desarrollar
- 6 Bibliografía

Objetivos Específicos

- Representar los mensajes de correo electrónico de forma estructurada.
- Extraer la información relevante de los contactos de correos electrónicos.
- Extraer conocimiento de los mensajes de correo electrónico y los archivos adjuntos.
- Desarrollar un prototipo de software que permita extraer, visualizar y analizar la información de los correos y las redes contenidas en ellos.

Metodología. I

- La primera fase del trabajo consiste en realizar una cuidadosa revisión bibliográfica de las técnicas de extracción de conocimiento en textos, específicamente en correos electrónicos.
- Revisión de algunas técnicas de gestión de conocimiento explícito y tácito, incluyendo la generación y análisis de redes sociales.

Metodología. II

- Luego de revisar las técnicas, se debe construir una representación estructurada, como XML, 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 ENRON¹.
- Extraer la información relevante de origen, destino y demás meta-datos, como, asunto, fecha, cuerpo y archivos adjuntos para realizar los análisis respectivos.

Metodología. III

- 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 diseñar e implementar un algoritmo de organización gráfica que me permita ver la red de esta forma, asignando además algún tipo de importancia al contacto y la conexión.

Metodología. IV

- Generar un prototipo que integre los desarrollos realizados durante las fases iniciales del proyecto, siguiendo una metodología orientada a prototipos. Se generarán cuatro 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 constuyen aparte, de modo que el sus procesos de desarrollo no interfieren. El tercer prototipo será la integración preliminar, y el cuarto, será el prototipo que integra todos los componentes de una forma estable.

¹<http://www.isi.edu/~adibi/Enron/Enron.htm>

Actividades a Desarrollar. I

1 Revisión Bibliográfica

- 1 Métodos de extracción de conocimiento de correos electrónicos
- 2 Gestión de Conocimiento tácito y explícito
- 3 Análisis de redes sociales

2 Escritura del Documento Final

- 1 Revisión y actualización del estado del arte (Durante todo el proyecto)
- 2 Diseño y documentación de la arquitectura de software
- 3 Descripción del prototipo
 - 1 Manuales

3 Desarrollo de Prototipos

- 1 Desarrollo del prototipo I

Actividades a Desarrollar. II

- ① Especificación (Entradas/Salidas)
- ② Arquitectura de software
- ③ Análisis de requerimientos
- ④ Diseño
- ⑤ Implementación
- ⑥ Pruebas

② Desarrollo del prototipo II

- ① Especificación (Entradas/Salidas)
- ② Arquitectura de software
- ③ Análisis de requerimientos
- ④ Diseño
- ⑤ Implementación
- ⑥ Pruebas

③ Desarrollo del prototipo III

Actividades a Desarrollar. III

- ① Arquitectura de software
 - ② Análisis de requerimientos
 - ③ Diseño
 - ④ Implementación
 - ⑤ Pruebas
-
- ④ Desarrollo del prototipo IV
 - ① Refinamiento del diseño
 - ② Pruebas globales de funcionamiento
 - ⑤ Documentación (Durante todo el proyecto)

Bibliografía Básica. I



Ulrik Brandes and T. Willhalm.

Visualization of bibliographic networks with a reshaped landscape metaphor.

In *VISSYM '02: Proceedings of the symposium on Data Visualisation 2002*, pages 159–ff, Aire-la-Ville, Switzerland, Switzerland, 2002. Eurographics Association.



Peter Busch, Debbie Richards, and C. N. G. 'Kit' Dampney.

The graphical interpretation of plausible tacit knowledge flows.

In *CRPITS '24: Proceedings of the Australian symposium on Information visualisation*, pages 37–46, Darlinghurst, Australia, Australia, 2003. Australian Computer Society, Inc.

Bibliografía Básica. II



Peter Anthony Busch, Debbie Richards, and C. N. G. 'Kit' Dampney.

Visual mapping of articulable tacit knowledge.

In *CRPITS '01: Australian symposium on Information visualisation*, pages 37–47, Darlinghurst, Australia, Australia, 2001. Australian Computer Society, Inc.



Soumen Chakrabarti.

Data mining for hypertext: a tutorial survey.

SIGKDD Explor. Newsl., 1(2):1–11, 2000.

Bibliografía Básica. III



Philipp Cimiano, Andreas Hotho, Gerd Stumme, and Julien Tane.

Conceptual knowledge processing with formal concept analysis and ontologies.

In Springer, editor, *Proceedings of the The Second International Conference on Formal Concept Analysis (ICFCA 04)*, pages 189 – 207, 2004.



Iain L. MacDonald and Walter Zucchini.

Hidden Markov and Other Models for Discrete-Valued Time Series.

Chapman & Hall/CRC, 2002.

Bibliografía Básica. IV



David Masterson and Nicholas 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.



I Nonaka and H Takeuchi.

The knowledge creating company.

Oxford University Press, 1995.

Bibliografía Básica. V



Fabien Picarougne, Gilles Venturini, and Christiane Guinot.

Un algorithme génétique parallèle pour la veille stratégique sur internet.

In *Veille Stratégique Scientifique et Technologique VSST 2004*, volume 2, pages 519 – 528, Toulouse, France, Octobre 25 – 29 2004.



Uta Priss.

A graphical interface for document retrieval based on formal concept analysis.

In Eugene Santos, editor, *Proceedings of the 8th Midwest Artificial Intelligence and Cognitive Science Conference*, pages

Bibliografía Básica. VI

66 – 70, Dayton Ohio, May 30 – June 1 1997. Wright State University.



Uta Priss.

Formal concept analysis in information science.

Annual Review of Information Science and Technology, ARIST,
40:521 – 543, 2006 (To be published).



II Richard J. Cole and Peter 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*, pages 309–315, London, UK, 1999. Springer-Verlag.

Bibliografía Básica. VII



Mehran Sahami, Susan Dumais, David Heckerman, and Eric Horvitz.

A bayesian approach to filtering junk E-mail.

In *Learning for Text Categorization: Papers from the 1998 Workshop*, Madison, Wisconsin, 1998. AAAI Technical Report WS-98-05.



Sunita Sarawagi and William W. Cohen.

Semi-markov conditional random fields for information extraction.

In *Advances in Neural Information Processing Systems*, December 13 2004.

Bibliografía Básica. VIII



Israel Spiegler.

Knowledge management: a new idea or a recycled concept?
Commun. AIS, 3(4es):2, 2000.



Jamie Twycross and Steve Cayzer.

An immune-based approach to document classification.
Technical Report HPL-2002-292, Information Infrastructure
Laboratory HP Laboratories Bristol, October 30 2002.