

Sistema para Identificar, Extraer y Visualizar Conocimiento en Correos Electrónicos Personales

Avance del Estado de la Tesis

Juan David Cruz Gómez
Laboratorio de Investigación en Sistemas Inteligentes
Departamento de Ingeniería de Sistemas e Industrial
Universidad Nacional de Colombia
Bogotá D.C., Colombia
jdcruzg@unal.edu.co

ABSTRACT

En este artículo se muestra el avance del estado del proyecto de creación de un sistema para extraer, gestionar y visualizar el conocimiento contenido en correos electrónicos personales. La idea es entonces extraer la información de los correos que hace referencia al origen y al destino para identificar las relaciones entre personas y luego identificar aquellos objetos que hacen que la relación identificada tenga sentido.

Keywords

Gestión de conocimiento, Análisis de redes sociales

1. INTRODUCCIÓN

En los últimos años, la gestión de conocimiento se ha convertido en una disciplina bastante difundida, ya que las empresas buscan manejar de alguna forma todo ese conocimiento que poseen sus empleados y así producir mayor valor en su compañía. Para realizar esta gestión, se han desarrollado múltiples técnicas, como el análisis de redes sociales. Adicionalmente, se han realizado diferentes estudios acerca de la forma en que el conocimiento es generado y posteriormente interiorizado en las personas. En este caso, se busca combinar el análisis de redes sociales con la gestión de conocimiento tácito, de modo que se pueda extraer la mayor cantidad posible de información –conocimiento– de los mensajes de correo electrónico que una persona envía y recibe. El artículo está organizado de la siguiente manera: en la sección 2, se define el problema y se describe la arquitectura propuesta para solucionarlo. EN la sección 3, se muestran algunos resultados preliminares obtenidos. Y, en la sección 4 se describen algunas conclusiones preliminares y se plantean las tareas globales a seguir.

2. DEFINICIÓN DEL PROBLEMA Y ARQUITECTURA PROPUESTA

2.1 Definición del Problema

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

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

2.2 Arquitectura Propuesta

Con el problema planteado, es necesario definir una arquitectura que permita localizar, pre-procesar y extraer la información buscada en los correos electrónicos. La localización, se refiere a la forma en que están almacenados los mensajes de correo, por ejemplo, en un servidor POP3¹ o en el formato de algún cliente de correo particular, como Microsoft Outlook o Mozilla Thunderbird. El pre-procesamiento se refiere a las transformaciones de cada uno de los mensajes de correo para facilitar su manejo y el proceso de extracción de características. La información buscada, está dividida en dos categorías, la primera, se refiere a toda la información que permite crear las redes sociales, esto es, remitente, destinatarios, fecha y asunto. La segunda categoría, es la que corresponde al objeto por el cuál el correo es enviado, es decir, la información que indica la razón por la que el mensaje se da: el objeto intermediador presente en el mensaje. El primer paso consiste entonces en la creación de un sistema que permita realizar las dos primeras tareas mencionadas. Para esto, se desarrolló la librería para correos electrónicos del LISI² de la cuál se muestra un esquema en la figura 1. Se ve entonces, que se pueden acceder diferentes orígenes de correos. Cada uno de los mensajes es convertido a XML

¹Post Office Protocol

²Laboratorio de Investigación en Sistemas Inteligentes de la Universidad Nacional de Colombia.

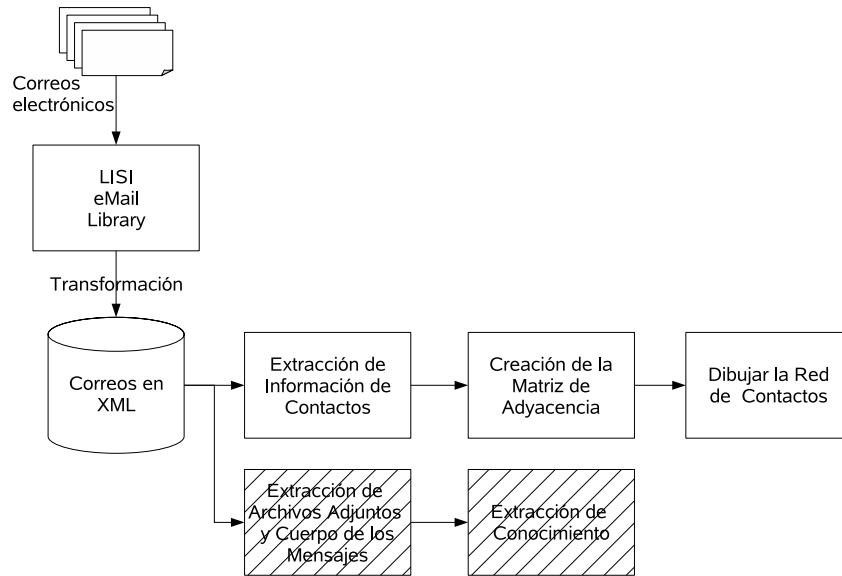


Figure 2: Esquema del sistema de extracción de conocimiento

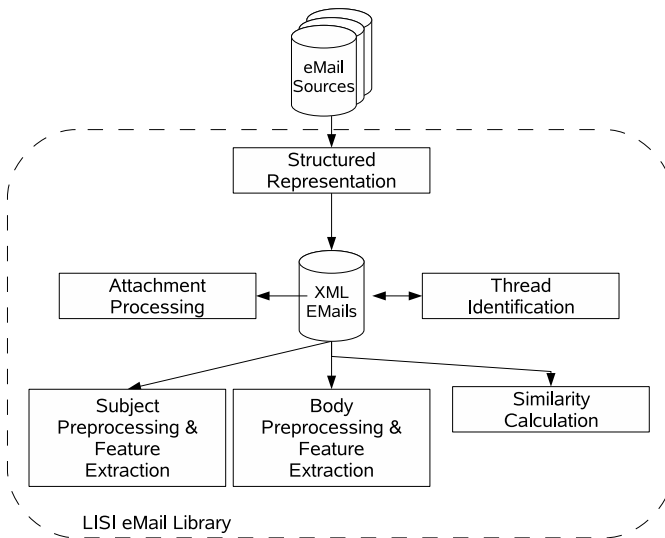


Figure 1: Esquema general de uncionamiento de la librería para correos electrónicos del LISI

utilizando XMTP, XML MIME Transformation Protocol, propuesto por Jonathan Borden [1]. Cada elemento en la representación XML corresponde a los campos propios del mensaje de correo electrónico, por ejemplo, *From*, *To*, *Subject*, entre otros. Cada uno de los mensajes es almacenado como un archivo, de esta forma ya están disponibles para ser utilizados por cualquier aplicación, como la de extracción de hilos de mensajes o el generador de redes sociales. En la figura 2 se muestra el esquema general de funcionamiento del sistema propuesto. En este momento, está desarrollado la parte de extracción de información de contactos y la generación de la red social. Aún falta desarrollar la parte que se encuentra sombreada, la cuál se refiere al procesamiento y extracción de información de los archivos adjuntos y del cuerpo del mensaje. Hasta ahora la librería de correos solo

se ha utilizado para realizar la conexión con los orígenes y para estructurar los mensajes.

3. RESULTADOS

Para realizar las pruebas de la arquitectura propuesta, se utilizó un conjunto de correos electrónicos de ENRON³, el cuál se hizo público al finalizar la investigación de fraude contra esta corporación. Corresponde a los correos de 150 usuarios, y en total tiene cerca de 500000 mensajes; se puede encontrar información detallada sobre este conjunto de datos en [2]. Para las pruebas preliminares se utilizó la carpeta de mensajes de *p-allen* (*Phillip Allen*). Esto no corresponde a ningún criterio particular, fue una elección aleatoria. En este caso, aunque los mensajes están en el formato del servidor POP3, están almacenados en forma de archivos de texto independientes, de forma local. Con estos archivos se procede a generar la matriz de adyacencia, con la cuál se generará posteriormente la red social. El primer ejemplo es mostrado en la figura 3. En esta se ve como el dueño de los mensajes de correo está hacia el centro de la red. Este tipo de redes, son conocidas como *egocéntricas*, y su principal característica es que la mayoría de los individuos están relacionados con un solo individuo. La disposición mostrada en la figura 3, que permite mostrar el egocentrismo de la red, no fue generada con algún algoritmo, sino que fue generada de forma aleatoria y luego corregida a mano. Esta no es la única forma de ver la red; se pueden utilizar otros algoritmos, como el que organiza la red como en la figura 4. En esta disposición se pueden ver más claramente las relaciones entre individuos. Hasta ahora, solo se ha identificado la red social de un individuo, sin embargo, no se han identificado los objetos de las relaciones encontradas, es decir, el por qué de la relación.

4. CONCLUSIONES Y TRABAJO FUTURO

El trabajo realizado hasta ahora ha consistido en el desarrollo de las herramientas necesarias para manipular los correos electrónicos y generar las redes sociales. En las redes

³<http://www.isi.edu/~adibi/Enron/Enron.htm>

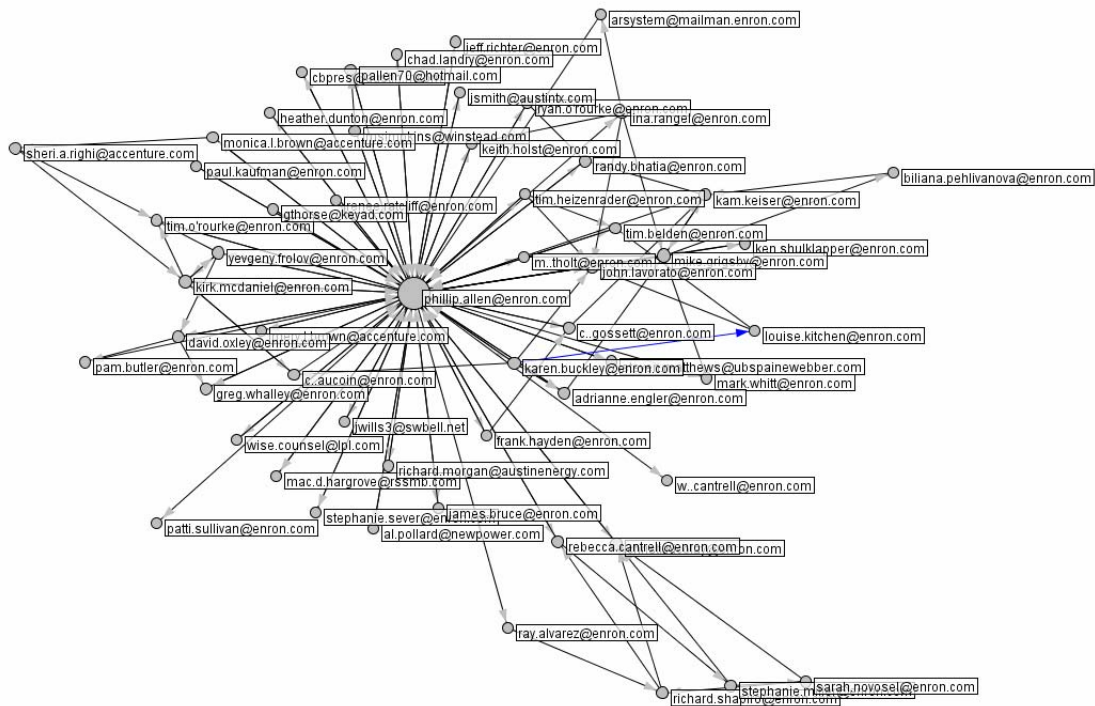


Figure 3: Ejemplo de la red egocéntrica de Phillip Allen. El tamaño de los nodos está dado por el grado que tienen.

generadas se puede ver algunas características importantes, y esperadas, como el alto grado que posee el dueño de los mensajes, así como el alto valor de la intermediación, es decir, el valor que mide que tan desconectada queda la red si este nodo desaparece.

En cuanto a las redes sociales, es necesario buscar un esquema de organización del grafo, de modo que es necesario estudiar otros algoritmos de organización que permitan además, asociar la información del cuerpo y de los mensajes adjuntos.

También queda por hacer, la extracción de información del cuerpo de los mensajes y de los archivos adjuntos, de modo que las relaciones sociales identificadas se puedan enriquecer.

5. REFERENCES

- [1] J. Borden. Xml mime transformation protocol (xmtp). Available in <http://www.openhealth.org/documents/xmtp.htm>, 2000. The OpenHealth Group.
- [2] B. Klimt and Y. Yang. Introducing the enron corpus. In *Proceedings of First Conference on Email and Anti-Spam CEAS 2004*, Mountain View, CA, July 2004.

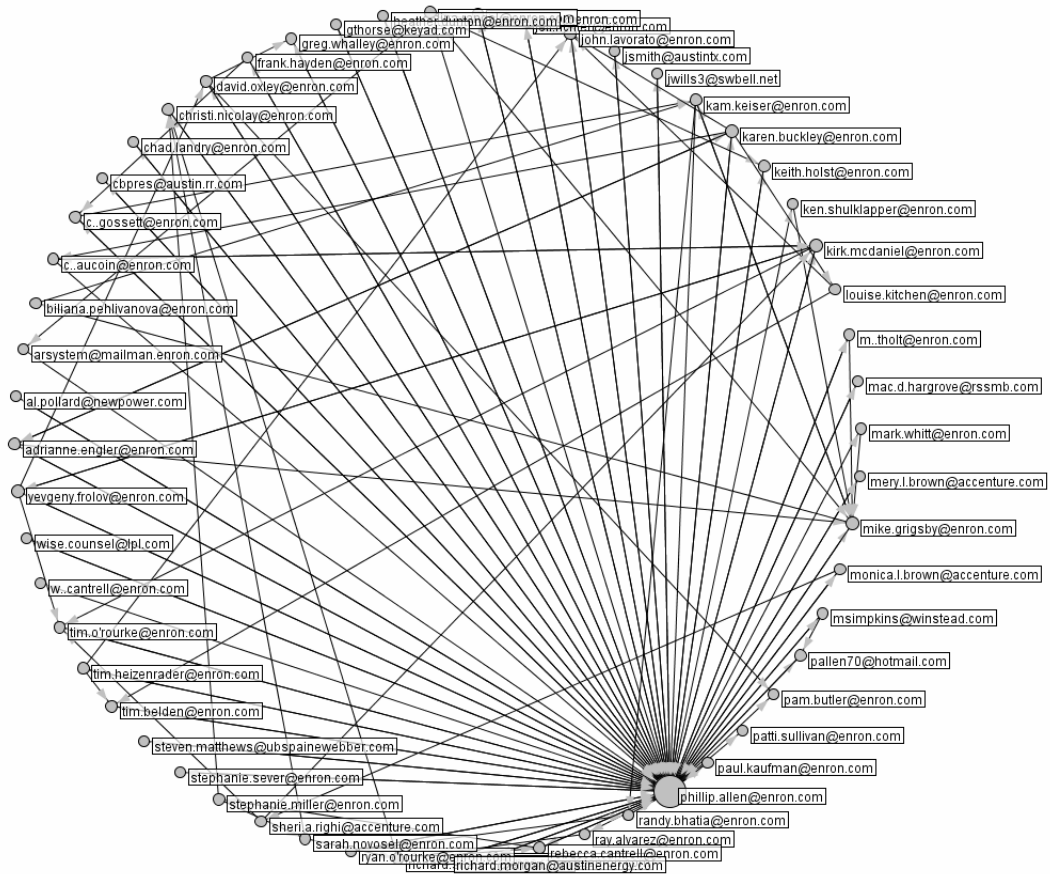


Figure 4: Ejemplo de la red circular de Phillip Allen. En este caso, el tamaño de los nodos también son diferentes.