1 Introducción

1.1 Motivación

En el mundo actual donde la información cambia de un día para otro, las redes sociales constituyen una de las expresiones del hombre para intercambiar y transferir lo que aprende y lo que crea. Dichas creaciones han permitido el acercamiento del público a la tecnología posibilitando la creación de verdaderas comunidades virtuales como es el caso de Facebook, Twitter o Linkedin, donde las personas se encuentran para discutir, relacionarse, intercambiar información y organizarse de forma relativamente similar a las comunidades presenciales. Según Nielsen Online, 67% de los usuarios de internet utilizan alguna de estas redes sociales para mantenerse en contacto con amigos, hacer crecer sus negocios o simplemente divertirse[1].

Twitter es quizá la más simple de las redes sociales porque consiste en enviar mensajes muy breves de 140 caracteres como máximo a aquellos que nos siguen (*followers*) y en recibir los mensajes de aquellos a los que seguimos; en esta simplicidad está su poder. No sirve para subir fotografías ni escribir una reflexión extensa, pero sí para que los usuarios informen a los demás de lo que hacen o viven en cada momento. Y es en este aspecto donde surgen algunas interrogantes: "¿Qué tiene de entretenido publicar lo que estoy haciendo?" o "¿por qué tienen otros que enterarse de las cosas que hago?". Y es que su poder va más allá, Twitter ha cambiado la forma de comunicarse llegando al punto en el que las noticias se dan primero por Twitter y luego por los medios convencionales. La novedad, que es el valor agregado que ofrecen las agencias de noticias, es ahora generada por los usuarios que desean participar y ser parte del nuevo proceso informativo, brindando información al instante con el uso de algún dispositivo móvil o alguna pequeña computadora.

Hay que aclarar que lo anterior no quiere decir que las redes sociales como Twitter desplazarán otros medios de comunicación masivos. No es lo mismo publicar unas líneas en Twitter, decir en unos cuantos segundos que algo está ocurriendo o mostrar las imágenes más espectaculares de un hecho, que explicar cómo y por qué sucede algo.

> Así, por ejemplo, el aviso en Twitter de un incendio alerta a quienes están conectados a la computadora. La transmisión en directo de la radio permite al radioescucha saber que una densa nube de humo se extiende en ese momento sobre determinada zona de la ciudad. Los televidentes podrán ver la dimen-

[1] CNBC 2010-09-10, "Social Networking: Your Key to Easy Credit? http://m.cnbc.com/us_news/34843251?refresh=true.

> sión de las llamaradas y testificar el valor de los bomberos que se internan en ese infierno. Al día siguiente, el lector de periódico podría conocer, por ejemplo, de qué manera los bomberos combatieron el incendio, las repercusiones que generarán las pérdidas, las historias de los empleados que perdieron su fuente de trabajo, los avances en las investigaciones sobre el origen del fuego, etcétera. Podría también leer una entrevista a fondo con uno de los bomberos o con una víctima, conocer si la ciudad cuenta con los recursos suficientes para atender siniestros de esa magnitud o saber qué hacer si se encuentra en una situación similar, entre muchas otras cosas[2].

Por último se puede decir que, con la posibilidad de llegar a más gente de manera inmediata (antes que los medios convencionales), el uso de las redes sociales ha revolucionado la necesidad de comunicarse y es que, aunque siempre hemos utilizado el potencial de las redes de forma intuitiva, la tecnología sólo ha acelerado el proceso de comunicación para no tener que decírselo a cada uno de todos nuestros conocidos. Y en este sentido Twitter expone los hechos que viven y reportan miles de usuarios en un solo momento.

1.2 Objetivos

El objetivo principal de este trabajo es desarrollar un sistema que permita el descubrimiento de información interesante en conjuntos de documentos extraídos de la red social Twitter.

Los objetivos específicos son:

- Implementación de bases de datos para la experimentación de la investigación.
- Evaluación experimental de la eficacia de modelos de espacios vectoriales como métodos para la representación de documentos.
- Implementación y evaluación de un algoritmo de agrupamiento aplicado sobre una de las bases de datos desarrolladas.
- Desarrollo de una aplicación empresarial web con Java EE que permita a usuarios la consulta de los conocimientos obtenidos por el sistema.

---

[2] Milenio online, 2010-05-09 ¿Para qué sirven los periodistas?
http://impreso.milenio.com/node/8764117. visitado el 11 de septiembre de 2010

Para lograr estos objetivos se siguió la siguiente estrategia y metodología:

a) Se realizó una revisión bibliográfica de los trabajos que se han desarrollado alrededor de la minería de textos en redes sociales.

b) Se llevó a cabo un análisis del modelo de representación vectorial de documentos y se establecieron sus características; así como de las medidas de distancia y similitud más importantes para la recuperación de información.

c) Se realizó una revisión de los métodos de agrupamiento jerárquicos y aglomerativos más importantes.

d) Una vez identificados los modelos más importantes para la representación de textos, las medidas de similitud y el algoritmo de agrupamiento, se elaboró una metodología para diseñar la arquitectura del sistema. El sistema se desarrolló en 4 subsistemas:

    a. Un subsistema para realizar consultas al servidor de Twitter con ayuda de Twitter4J, una librería escrita en el lenguaje de programación Java de código libre que implementa la API de Twitter. El mismo subsistema se encarga del pre-procesamiento del texto y la inclusión o no de ese mensaje de Twitter a la base de datos, consiguiendo así una clasificación.

    b. El segundo sistema es responsable de la representación vectorial y agrupamiento de los documentos.

    c. Un tercer subsistema se encarga de recibir los grupos generados por el anterior subsistema, almacenarlos en una base de datos y presentar dicha información.
    d. Y, finalmente, un cuarto encargado de modificar la base de datos de los grupos generados anteriormente.

1.3 Límites de la tesis

La red social Twitter por su propia naturaleza contiene una gran cantidad de información difícil de manejar. Por ello, una limitación natural del estudio es la imposibilidad de manejar toda la información que se genera, por lo que se enfocan los esfuerzos a una determinada zona geográfica y la recopilación de temas definidos a priori.

El presente trabajo no pretende ser un nuevo paradigma computacional para el desarrollo de sistemas de minería de textos, pues solo propone una aplicación de dichos sistemas a la red social Twitter.
No se presenta un método de evaluación para el algoritmo de agrupamiento, ya que los métodos existentes evalúan la eficacia con respecto al valor de la distancia de corte del dendograma, las muestras asignadas a un grupo y muestras que no deberían estar dentro de un grupo, cosa que resulta particularmente difícil puesto que el sistema recibe un conjunto variable de gorjeos y el dendograma para cada conjunto de datos varía para cada conjunto.

Por el momento el sistema solo presenta a los usuarios el resultado del agrupamiento de los gorjeos, no es posible hacer consultas personalizadas y no se implementan métodos de resumen automático; aspectos que se pretenden cubrir en trabajo futo.

1.4 Vista general de la tesis

En el capítulo 2 se hace una breve reseña sobre los temas tratados en el presente trabajo, temas como son la minería de textos, la minería web y las redes sociales.

El capítulo 3 versa sobre el modelo de representación vectorial; un modelo algebraico muy utilizado en el área de recuperación de información, donde se representan documentos del lenguaje natural por medio de vectores en un espacio lineal multidimensional.

El capítulo 4 hace una introducción al tema de agrupamiento o *clustering*, medidas de similitud y distancia para finalmente introducir al algoritmo de agrupamiento jerárquico utilizado en esta tesis. Describiendo de esta manera los elementos teóricos necesarios para comprender la propuesta de solución al problema de agrupamiento que se presenta.

El capítulo 5 introduce al lector a la tecnología de Java empresarial, fundamental para el desarrollo del presente trabajo por ofrecer los recursos tecnológicos para la óptima implementación del sistema.

El capítulo 6 detalla la implementación del sistema; la consulta a la base de datos de Twitter en tiempo real, la caracterización de entrada al algoritmo de agrupamiento y los pasos para conseguir grupos que reflejen relaciones semánticas.

El capítulo 7 trata del análisis de resultados; cualitativo y cuantitativo de los grupos generados por el sistema.

Finalmente dentro del capítulo 8 se hace un resumen de la tesis y se presentan las aéreas de oportunidad para trabajos futuros.