Clasificación de Tweets Utilizando el Modelo Probabilístico de Tópicos Laten Dirichlet Allocation

Dr. Francisco Jacob Ávila Camacho*, M en ADN. Juan Manuel Stein Carrillo**

Acerca de los autores...

Div. de Sistemas Computacionalés, Tecnológico de Estudios Superiores de Ecatepec
fjacobavila@tese.edu.mx
jmsteinc@tese.edu.mx
*, **Académicos de la División de Ingeniería en Sistemas Computacionales del Tecnológico de Estudios Superiores de Ecatepec.

Resumen

Twitter[1] es uno de los servicios de microblogging que se ha vuelto muy popular dentro de los medios sociales utilizándose como una herramienta de comunicación. Las enormes cantidades de entradas, conocidas como tweets, que se generan todos los días, no solo representan un problema en el manejo de la información, sino un tema de investigación interesante, dado que representan una vasta fuente de información con interés para la minería y el descubrimiento de conocimiento. Los usuarios de Twitter están conectados por relaciones de "seguimiento", es decir, una persona obtiene tweets por seguir a otra persona, de esta manera podemos construir una red de información asociada por texto para un mejor modelado y el descubrimiento de patrones de interés acerca de los datos de texto. Con este trabajo se desarrollan tres tareas de investigación sobre los datos de texto en Twitter: el filtrado de tweets, basado en los intereses de un usuario; la creación de comunidades de un grupo grande de personas, y la clasificación de tweets. Identificando intereses, podemos filtrar informaciones no deseadas de los tweets de entrada, la creación de comunidad ayuda a encontrar subgrupos de intereses en particular y sugerir usuarios con intereses similares para seguirlos, y la clasificación de tweets ayudará a los usuarios a seleccionar sus categorías de tweets favoritas para leerlas. Los casos experimentales son diseñados y ejecutados para demostrar la efectividad de la plataforma y los algoritmos estadísticos propuestos para estas tareas.

Palabras clave

Minería de datos, redes sociales, descubrimiento de conocimiento, microblog, clasificación de tweets.

Abstract

Twitter[1] is one of the microblogging services that has become very popular within social media using it as a communication tool. The huge numbers of posts, known as tweets, which are generated every day, represent not only a problem in the handling of information, but an interesting research topic, as they represent a vast source of information with interest to the mining and knowledge discovery. Twitter users are connected by "tracking" relationships, i.e. one person gets tweets by following someone else, so we can build a network of associated information by text for better modeling and the discovery of patterns of interest in text data. This work develops three research tasks on text data on Twitter: tweet filtering, based on a user's interests; creating communities for a large group of people, and ranking tweets. By identifying interests, we can filter unwanted information from input tweets, creating community helps to find particular subgroups of

interests and suggest users with similar preferences to follow them, and ranking tweets will help users to select their favorite tweet categories to read. Experimental cases are designed and executed to demonstrate the effectiveness of the proposed platform and statistical algorithms for these tasks.

Keywords

Data mining, social media, knowledge discovery, microblog, tweet classification.

1. Introducción al Microblog

El microblog es una herramienta de comunicación en línea a través de la cual los usuarios indican lo que están haciendo o pensando en el momento, cuál es su opinión sobre un tema o fenómeno en específico, establecer una conversación tipo chat, compartir información, comentar sobre noticias, así como llevar a cabo proselitismo político.

El microblgging se ha hecho muy popular en la red, ya que permite a los usuarios propagar actualizaciones de textos breves hacia un público o a un grupo limitado de contactos.

Las entradas de microblog tienen muchas características diferentes a los documentos tradicionales de texto, por ejemplo en Twitter, la longitud máxima de una entrada de microblog, conocida como tweet, es de 140 caracteres. Dado que los mensajes son muy cortos, existe una fuerte habilidad para expresar ideas de manera efectiva y compartir información como una herramienta de comunicación en la red.

1.1 Twitter como una fuente de datos

Los datos en Twitter crecen extremadamente rápido. De acuerdo con estadísticas no oficiales, existen alrededor de 155 millones de tweets publicados en la red todos los días. Esto es más de tres veces los 50 millones de tweets por día que reportó oficialmente la empresa hace un año. Adicionalmente a la enorme cantidad de datos de texto, la gente que utiliza Twitter está interconectada por un tipo de relación especial: el seguimiento. Seguir a una persona significa que alguien está interesado en sus tweets, por lo que esta, se alimenta de los tweets.

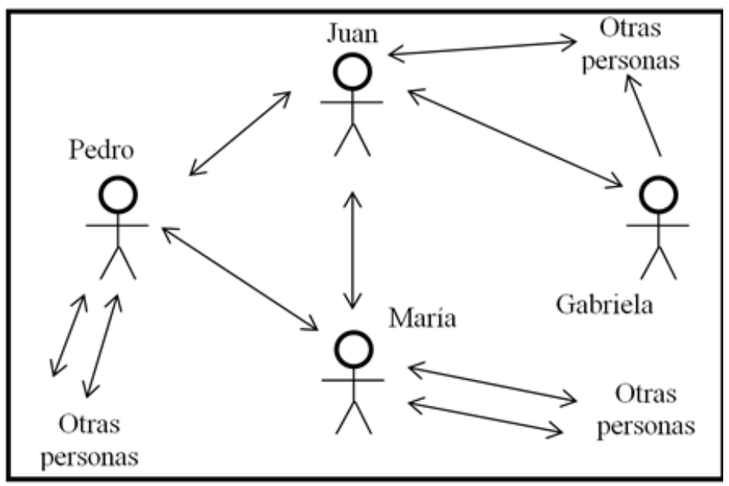


Figura 1

Relación de “Seguimiento” en Twitter que puede ser utilizada para crear una red de información.

La Figura 1 muestra un esquema de relación de seguimiento mutuo. Bajo estas relaciones de seguimiento, los usuarios de Twitter y sus tweets asociados forman una red de información creciente. Cada tweet de un usuario es un nodo y cada relación de seguimiento es un vértice en la red de información, respectivamente. Por ejemplo, una línea bidireccional entre Juan y Pedro indica un vértice entre dos nodos creados por sus tweets. De esta forma, Twitter se convierte en un conjunto de datos único y prometedor para descubrir patrones interesantes. Tanto los objetivos de la minería en texto, como la administración de la información y la búsqueda de temas en microblog, son temas interesantes de investigación por estar estrechamente relacionados a las características especiales de los tweets.

Algunas particularidades de los datos de Twitter, como la corta longitud de los mensajes, las consideraciones de tiempo, el uso de las etiquetas hash (hash tags) y la interconexión entre usuarios, representan retos como el conocimiento previo en descubrimiento de patrones. Adicionalmente, dado que Twitter ofrece una API para obtener los tweets sin costo, se utilizó de manera natural como fuente de datos para desarrollar los experimentos de este trabajo.

1.2 Tareas de investigación

1.2.1 Filtrado de Tweets basado en los intereses del usuario

Dada la definición proporcionada por Wikipedia[4], un sistema de filtrado de información es aquél que elimina información redundante o no deseada proveniente de un flujo de información, utilizando métodos automatizados o semi-automatizados antes de ser proporcionado al usuario. Esta definición puede ser aplicada al filtrado de información en Twitter: antes de que un flujo de tweets sea entregado al usuario, idealmente existe un sistema para filtrar los tweets no deseados y obtener únicamente los de interés para el usuario.

Es una escena común que cuando una persona está siguiendo a otras que publican tweets activamente, cada día se encuentra con una enorme cantidad de ellos, pero no siempre se está interesado en todos y sólo se desea recoger algunos para leerlos.

Este trabajo propone un método basado en un umbral para filtrar tweets sin interés. El interés de un tweet está dado por un valor numérico. Si este valor es mayor a un umbral dado, éstos serán retenidos y presentados al usuario, de otra manera, serán filtrados antes de que le sean mostrados.

El indicador de interés será calculado vía la probabilidad marginal de un tweet, dado un modelo que representa los intereses de los usuarios. Este modelo será estimado por medio de la Ubicación Latente de Dirichlet (Latent Dirichlet Allocation – LDA) [5, 14, 16], un modelo probabilístico de tópicos que representa una colección de documentos como un conjunto de distribución de tópicos para cada uno. Los intereses de las personas son los tópicos en el método LDA, y el interés de un usuario se representa por una distribución de tópico de interés.

1.2.2 Descubrimiento de Comunidad

En las redes sociales como Facebook y Twitter, la tarea de formar una comunidad significa que un determinado número de personas crea subgrupos entre quienes son “similares” dentro del grupo. La medida de similitud entre personas significa para este trabajo la similitud de intereses entre ellas. Una persona en particular, puede o no estar involucrada en uno o varios subgrupos al mismo tiempo, lo que dependerá de las necesidades de la aplicación hecha por el usuario.

Esta funcionalidad de proporcionar a las personas, recomendaciones de usuarios similares es casi indispensable para la aplicación, ya que beneficia, pero no limita, a los siguientes aspectos: ayuda a encontrar otras personas con las que pudiera haber interés para seguirlas. También ayuda a obtener más usuarios conectados y estimula el crecimiento de su red social. Adicionalmente, proporciona una forma posible de crear publicidad en línea, basado en los intereses comunes del subgrupo.

El problema básico de investigación es encontrar comunidades (subgrupos) de usuarios con intereses similares. Adicionalmente a la explotación de los textos de información generados por los tweets, existen conexiones, las relaciones de seguimiento entre las personas. Si una persona está siguiendo a otra, es muy probable que compartan algún interés. Como resultado, la información de conexión puede ser explotada adicionalmente a la información de texto para crear un mejor modelo de tópicos. La idea fue obtenida del artículo de Yizhou "Introducing to iTopicModel [6]" en la cual se utiliza información de texto y de enlaces entre documentos dentro de la red de información para mejorar el modelo tradicional de tópicos. Mediante las relaciones de seguimiento en Twitter, podemos construir una red de documentos. Después de que se construye el modelo de tópicos, al comparar la distribución de tópicos de los usuarios, podemos encontrar subgrupos que comparten los mismos intereses y eso ofrecerá en las recomendaciones de grupos.

El subgrupo se define como un conjunto de nodos en la red de información experimental, donde el número de nodos es mayor a un umbral dado y la similitud entre pares de ellos es mayor a otro umbral dado. A lo mucho, la distancia entre dos nodos que son similares en intereses de dos usuarios, se define como una similitud coseno de las dos respectivas distribuciones de tópicos.

1.2.3 Clasificación de Tweets

El objetivo de la clasificación de tweets es dividir diferentes tweets en categorías, dependiendo del contenido de cada uno. El propósito de esto es que a menudo los usuarios de Twitter desearían seleccionar unos cuantos tweets sobre categorías específicas para leerlos en algunos momentos específicos.

Por ejemplo, ciertos profesionistas preferirían leer algunos tweets técnicos mientras trabajan, ver determinadas noticias para relajarse, algunos twees de conversación en tiempos de ocio y observar lo que hacen las personas en las que tienen interés. Ofreciendo varias categorías con tweets relacionados, los tweets de un usuario se podrían mostrar en varias columnas, permitiéndole seleccionarlas para leerlas en diferentes situaciones.

Para lograr dicha tarea, se aplicará y modificará el algoritmo K-Means [9, 10, 17], que es sencillo pero efectivo. Para clasificar un número de tweets dentro de algunas categorías comunes, como discusiones técnicas, noticias, eventos, deportes, conversación diaria, lo cual representa el comportamiento habitual de los usuarios, se modificará el algoritmo K-Means especificando manualmente algunos tópicos particulares, para que de esta manera, el algoritmo clasifique los tweets y recalcule los centros de una forma iterativa. Con la ayuda de supervisión, los resultados de clasificación pueden ser mejor ejecutados de acuerdo con las necesidades del usuario.

2. Modelado

2.1 Filtrado de tweets basado en los intereses del usuario

En esta sección, se proporciona una definición formal y el modelo probabilístico para filtrar los tweets carentes de interés, dependiendo de ciertos criterios del usuario. Como se mencionó anteriormente, para determinar si un tweet es de interés o no para el usuario, si debe ser retenido o filtrado, se debe aplicar un modelo de tópicos de interés, construido con el modelo LDA propuesto por David Blei [5], para estimar el modelo que representa los intereses del mismo, y así calcular la probabilidad marginal de un tweet como un indicador de interés.

LDA como un modelo generativo, utiliza una distribución multinomial sobre tópicos controlados por un Dirichlet previo para un documento. La función de la densidad de probabilidad de una distribución Dirichlet sobre una distribución multinomial p, está dada por:

$$p(\theta|\alpha) = \frac{\tau(\sum_{i=1}^{k}\alpha_i)}{\prod_{i=1}^{k}\tau(\alpha_i)}\theta_1^{\alpha_1-1}\theta_2^{\alpha_2-1}\dots\theta_k^{\alpha_k-1}$$

Donde α es un parámetro a nivel de colección, el vector de longitud k como hiper parámetro para la distribución Dirichlet, donde k es el número de tópicos en la colección. τ es la función beta[].θ es un parámetro a nivel de documento, el vector de longitud k indica la distribución de tópicos de un documento. En nuestro caso, la colección de documentos es un cúmulo de tweets adquirida de personas que el usuario está siguiendo en Twitter.

El proceso generativo para un tweet *t* en el modelo, se puede describir de la siguiente forma:

1) Seleccionar la longitud del tweet N~ Poisson(⍰)

2) Seleccionar la distribución de tópicos θ~Dir(α) para *t*

3) Para cada una de las N palabras w en el tweet,

a. Seleccionar un tópico z ~Multinomial(θ)

b. Generar una palabra w bajo p(w|z,β), una probabilidad multinomial condicionada al tópico.

Bajo el modelo LDA, la probabilidad marginal de un nuevo tweet *t* de entrada está dada por:

$$p(t|\alpha,\beta) = \int_{\alpha} p(\theta|\alpha)\prod_{j=1}^{N}p(w_j|t) = \int_{\alpha} p(\theta|\alpha)(\prod_{j=1}^{N}\sum_{k=1}^{k}p(z=k|\theta)p(w_j|z=k,\beta))d\theta$$

El logaritmo de esta probabilidad se define como el indicador de interés utilizado para compararlo contra el umbral λ de acuerdo con la condición de filtrado.

$$\text{score}(t,\alpha,\beta)=\log⟦p(t|\alpha,\beta)⟧$$

Si score $(t,\alpha,\beta)>\lambda$, este tweet *t* es definido como "interesante" al usuario, por lo que será retenido. De lo contrario, es "no interesante" y por lo tanto será filtrado antes de presentárselo.

De esta forma, el proceso completo de filtrado de tweets para un usuario en particular, se describe de la siguiente manera:

1) Obtener un número de tweets de las personas que actualmente sigue un usuario.
2) Tratar estos tweets como documentos para estimar el modelo LDA como el modelo de interés del usuario.
3) Dado un flujo de nuevos tweets de entrada, para cada tweet calcular la probabilidad marginal bajo el modelo estimado de intereses y comparar con el umbral de filtrado para decidir si el tweet debe ser filtrado o no.
4) Presentar los tweets retenidos en el paso anterior, los cuales son considerados como interesantes al satisfacer las necesidades de información del usuario.

2.2 Creación de comunidad

Para encontrar comunidades entre un grupo grande de personas, una forma es encontrar subgrupos que comparten intereses similares. El interés de una persona puede ser modelado como una distribución de tópico sobre todos los tópicos de la colección, la cual está compuesta por tweets provenientes de todas las personas bajo el experimento.

La distribución de tópicos de interés para cada persona puede ser estimada utilizando la técnica de modelado de tópicos. Adicionalmente, son considerados los enlaces entre ellas, los cuales representan la relación de seguimiento.

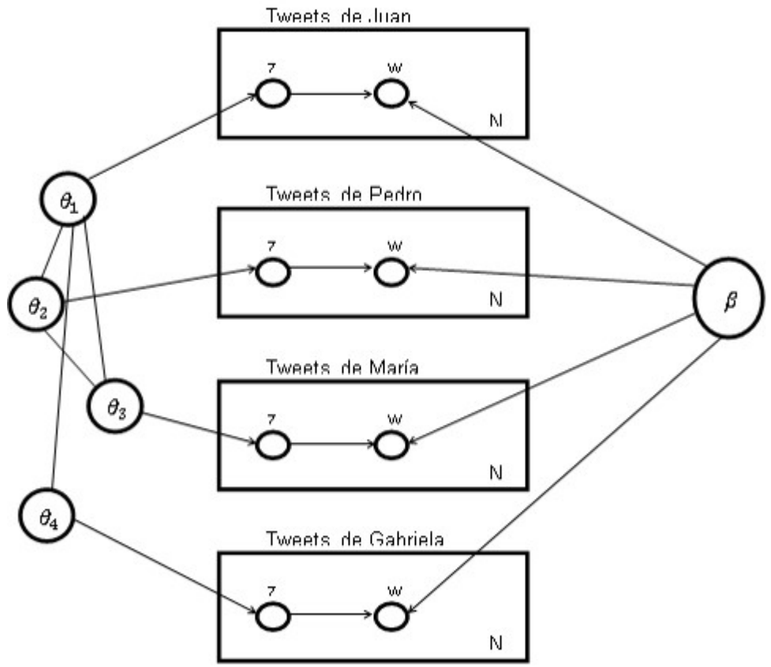


Figura 2

Modelo de tópicos en la red de información que explora texto y enlaces.

La Figura 2 es la representación del nuevo modelo gráfico para la red de información de Twitter, seguida por el modelo iTopicModel [6]. Cada θ_i es una distribución de tópicos de los tweets de un usuario. Cada rectángulo es el proceso de generación de los tweets respectivos. El modelo es diferente del tradicional de tópicos en el lado izquierdo de la figura: no hay dependencias entre aquellas distribuciones de tópicos del modelado tradicional. Las dependencias están directamente relacionadas y modeladas por la relación de seguimiento mutuo en la red de información de Twiiter. La suposición es que las personas que se siguen mutuamente, es porque comparten intereses similares.

Para nuestro trabajo, podemos agrupar 1,000 tweets de un usuario como un documento, y si dos usuarios se siguen mutuamente, habrá un enlace entre sus distribuciones de tópicos y éstas eventualmente afectarán a cada uno.

Los enlaces reforzarán a sus vecinos por sus conocimientos, y la información propagada podrá generar mejores modelos.

Estos enlaces en el modelo gráfico son modelados por campos aleatorios de Markov [11], que proporcionan probabilidad estructural, i.e. p(θ_ |G). Específicamente,

$$p(\theta \ \ |G) = \frac{1}{Z}\exp\{-\sum_{c\in C} Vc(\theta)\}$$

Donde C es un conjunto de grupos en el gráfico G,

$$Z = \sum_{\theta} \exp\{-\sum_{c\in C} V_c(\theta)\} \ es\ una\ función\ de\ partición$$

Y Vi(θ) es una función potencial definida como:

$$V_i(\theta_i) = -(\alpha_i^0 - 1)^T \log(\theta_i)$$

$$V_{i\to j}(\theta_i, \theta_j) = -(w_{ij}\theta_j)^T \log(\theta_i), Si < t_i, t_j > \in E;$$

De lo contrario, es igual a 0, si el usuario i está siguiendo al usuario *j* hay un vértice en la gráfica y *w_ij*=1 indicando el peso del vértice.

De esta manera, la probabilidad estructural se deriva de:

$$p(\theta|G) = \frac{1}{Z} exo\{\sum_{i}\left[\left(\alpha_i^0 + \sum_{j\in N(i)} w_{ij}\theta_i - 1\right)^T \log(\theta_i)\right]\}$$

Al mismo tiempo, el modelo tradicional de tópicos proporciona la probabilidad de los tweets, *i.e. p(t_i |θ,β)* para cada tweet. Específicamente, los tweets agrupados de un usuario denotados como ti es estimado como un N –documento de palabras en el modelado de tópicos y también modelado por una mezcla de modelos sobre K tópicos, donde cada documento se asume que es independiente de los demás.

Siendo así, la probabilidad de la colección completa T para M –tweets de usuario, está dada por:

$$p(T|\theta,\beta) = \prod_{i=1}^{M} p(t_i|\theta,\beta)$$
$$= \prod_{i-1}^{M}\prod_{j-1}^{N} p(w_j|t_i,\theta,\beta)$$
$$= \prod_{i=1}^{M}\prod_{j=1}^{N}\sum_{k=1}^{K} p(z=k|x_i)p(w_j|z=k) = \prod_{i=1}^{M}\prod_{j=1}^{N}\sum_{k=1}^{K}\theta_{ik}\beta_{kj}$$

Con la probabilidad de estructura y textual, podemos tomar la probabilidad de unión como una función objetivo, y utilizar el algoritmo EM para estimar los parámetros del modelo, θ, el cual es un conjunto de distribuciones de tópicos para cada conjunto de tweets agrupados para cada usuario, y β que contiene la probabilidad de cada palabra bajo cada tópico.

Una vez que θ es estimado, se puede recolectar un conjunto de comunidades, dada la siguiente definición de comunidad,

comunidad=$\{c|\forall\theta\_(1,)\ \theta\_2\in c, c(\theta, similitud(\theta\_1,\theta\_2)> \varphi, |c|>\mu\}$

Donde μ es el número mínimo de usuarios en una comunidad y φ es la similitud mínima entre cualquier par de elementos dentro de ésta. La similitud coseno [12, 13] es seleccionada como la medida de similitud,

$$similitud(\theta_1,\theta_2) = \frac{\theta_1 * \theta_2}{||\theta_1||\theta_2||} = \frac{\sum_{i=1}^{K}\theta_{1j}X\theta_{2j}}{\sqrt{\sum_{j=1}^{K}\theta_{1j}^2}\,X\,\sqrt{\sum_{j=1}^{K}\theta_{2j}^2}}$$

Donde θ_1j es la probabilidad del jth tópico en la ith distribución de tópico de interés del usuario y K es el número total de tópicos, así como también la longitud de cada θ_i.

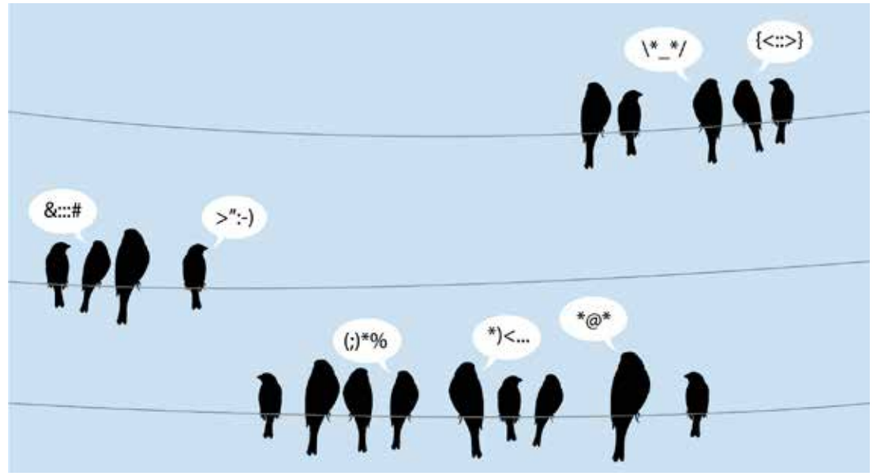
2.3 Clasificación de tweets

La tarea de clasificación de tweets se lleva a cabo modificando un poco el algoritmo K-Means: se crean manualmente un pequeño número de tweets de prueba para utilizarse como indicador de clasificación inicial, en lugar de obtenerlos aleatoriamente, que es lo que hace el algoritmo original.

Los tweets de prueba se crean para alguna categoría común de interés como discusión profesional, noticias, deportes o conversación diaria; por supuesto que estas categorías podrían cambiar de acuerdo con los intereses y preferencias del usuario.

Los tweets serán clasificados con base en la cercanía al indicador actual de manera iterativa. La medida de distancia entre dos tweets *p* y *q* seleccionados, es la distancia euclidiana dada por:

$$d(p,q) = \sqrt{(c_{11}-c_{21})^2+(c_{12}-c_{22})^2+\cdots+(c_{1n}-c_{2n})^2} = \sqrt{\sum_{i=1}^{|V|}(c_{1i}-c_{2i})^2}$$



Aquí, cada tweet se representa por un vector de conteo de palabras dado un vocabulario *V* que guarda cada palabra.

c1i y c2i son los conteos de las palabras en *p* y *q* respectivamente, donde *i* es la posición en la que se encuentra la palabra dentro del vocabulario.

El algoritmo se ejecuta de una forma iterativa y en cada iteración se clasifican los tweets de acuerdo con la cercanía con el indicador de cada categoría, para después recalcular nuevos indicadores para los tweets que ingresaron a una categoría, el proceso completo se describe con los siguientes pasos:

1) Transformar cada tweet para ser clasificado dentro de los vectores de conteo de palabras sobre el vocabulario.
2) Establecer el indicador inicial creando manualmente tweets de prueba.
3) Iterar sobre cada tweet para determinar la cercanía con cada indicador, comparando la distancia entre él y el indicador inicial, y así clasificarlo dentro de la categoría correspondiente.
4) Recalcular los indicadores de acuerdo con el indicador promedio de cada categoría.
5) Repetir los pasos (3) y (4) hasta que exista un mínimo cambio en la clasificación o se alcance el número máximo de iteraciones.

Después de que el paso 5 sea completado, se podrán mostrar los resultados de la clasificación al usuario por medio de columnas de tweets representando diferentes categorías.

3. Experimentos

3.1 Recolección de datos

Twitter proporciona una serie de interfaces de programación de aplicaciones (API) que permiten obtener o recuperar los tweets, datos de perfiles, actividades de usuarios o eventos de personas, así como la conexión que existe entre usuarios, es decir, las relaciones de seguimiento.

En la sección de desarrolladores de Twitter podemos obtener los detalles de las APIs, mismas que están basadas en el protocolo HTTP, por lo que se pueden utilizar para realizar peticiones de tipo GET o POST para acceder a los datos. La API requiere de una autenticación básica y tener una cuenta válida en Twitter; los datos se obtienen en formato XML o JSON.

En el experimento se utilizaron tres APIs: una que proporciona los identificadores IDs de las personas que siguen un usuario dado, otra que proporciona los 200 tweets más recientes del usuario dado y otra más que juzga si existe relación de seguimiento entre dos usuarios.

Para obtener mejores resultados, todos los datos obtenidos con la API fueron pre-procesados, este pre-procesamiento consistió en eliminar palabras auxiliares y de artículo como: a, the, that, etcétera en idioma inglés, para normalizar las palabras de manera uniforme.

3.2 Diseño del experimento y resultados

3.2.1 Filtrado de tweets con base en los intereses del usuario

La persona seleccionada para este experimento fue Hilary Mason **(http://twitter.com/hmanson)**, quien es jefa de investigación en bit.ly [7], y se encuentra interesada en el tema de la máquina de aprendizaje. Utilizando la API, se obtuvieron los 200 tweets más recientes de 150 personas que ella sigue, seleccionados al azar. El umbral de interés se estableció en -100.

De inicio se ingresaron 30 tweets nuevos al filtro, con el fin de establecer las categorías de interés. La Tabla 1 muestra únicamente 10 de los 30 tweets nuevos recibidos junto

Tabla 1
Flujo de 10 tweets de entrada, su probabilidad marginal y el resultado de filtrado.

Tweets de entrada nuevos	$\log(t\|\alpha,\beta)$	Estatus
New blog: What should I cut from Team Time Management?: I am rewriting my class Advanced Time Management	-138.08359	Filtrado
RT @Algebra: The condition number of a matrix A relative to the euclidian norm is the ratio of its smallest igenvalues	-107.43564	Filtrado
@gruber: how exactly are Mobile Safari exploits worse with Nitro? Remote address book hijack posible pre-Nitro	-99.891827	Retenido
Pretty interesting RT @OpenHQR: San Francisco Rainwater: Radiation 181 Times Above US Drinking Water Standard	-146.24928	Filtrado
Harley to slow for city traffic. This is a modified R1200C, also a massive machine but more torque for city driving	-100.24776	Filtrado
Reading #OSCON Data proposals, have coined a new acronym. YAWNS – Yet Another Wanking NoSQL Solution	-90.669528	Retenido
Your infographic has one design flaw – I impulsively want to hover over the points in the scatterplot and see the couples	-83.292602	Retenido
You are woking on exciting stuff that will revolutionize fashion Fashionistas and entrepreneurs stay tuned	-109.35694	Filtrado
Having a lot of fun with the Beads Processing Library, Easy, full-features sound synthesis, analysis, and playback	-111.00092	Filtrado
#followfriday @aghose NYU Stern professor, new Tweep, and one of the winners of the WWW2011 best paper award	-98.841583	Retenido

con su indicador de interés, el cual se obtuvo con el logaritmo de la probabilidad marginal de cada tweet dado por el modelo de interés que fue creado. Los tweets cuyo indicador de interés sea mayor a -100 son retenidos y el resto son considerados sin interés para el usuario. Dentro de la Tabla 1 se observa que los tópicos de los cuatro tweets retenidos están relacionados al tema de ciencias y el web, cuyos tópicos le preocupan al usuario. Los seis tweets filtrados hablan poco de ciencia y de la máquina de aprendizaje, que son tópicos de interés para el usuario.

3.2.2 Creación de comunidad

Para encontrar comunidades en una red de información, el primer paso es construir la red de información en Twitter con texto como el principal atributo para los nodos. Cada nodo de la red de información se crea agrupando 1,000 tweets de un usuario (algunos de ellos han publicado al momento menos de 1,000 tweets), el vértice entre dos nodos se forma si los dos usuarios están asociados con la relación de seguimiento mutuo.

Con el fin de proporcionar una forma de evaluar el resultado final, se seleccionaron 30 usuarios como nodos semilla y cuyos intereses son claros, para una búsqueda inicial amplia y de esta forma ser incorporados más usuarios como nodos a la red de información. El interés claro se determina buscando manualmente en sus tweets, checando si esos tweets son buenos indicadores para ciertos intereses. En el experimento, el interés claro incluye “cooking”, “Internet and web”, “multimedia”, “social networking”, “current affairs”, “digital library”, “travel”, “game industry” y más. Como evaluación comparamos los intereses de las comunidades encontradas con el modelo de tópicos integrado con los usuarios seleccionados previamente.

Después de recolectar 2’313,276 tweets de 400 usuarios como nodos, se agregaron los vértices entre dos nodos si los usuarios se seguían mutuamente. El modelo iTopicModel

Tabla 2
Comunidades creadas y su información relacionada.

	Comunidad 1	Comunidad 2	Comunidad 3	Comunidad 4
Número de usuarios	48	67	91	42
Valor mínimo de similitud	0.76	0.82	0.90	0.71
Tema de la comunidad	Game, Media, Entertainment	Travel, Events, World	Internet, Web, Online	Reaserch, Study
Principales Palabras clave	Games, games, startup, microtask, samplereality, gameloft, cpuo, ps3, love, crowd, play, flash, ea, starcraft2, ibogost, xbox, mobile, amazing, tv, free, experience, video, zynga	Tonight, pm, airport, waiting, Honolulu, hotel, international, Boston, checked, car, center, blog, Hawaii, world, interesting, article, photos, event, Libya, story, Egypt, piece, east, missing, chinese	http, mobile, live, space, tech, elearning, action, marketing, twitter, facebook, tweet, fb, follow, page, users, link, google, book, free, ipad, email, read, search, phone, cool, books, code, site, apps	Digital, library, university, research, culture, job, pdf, humanities, projects, conference, public, tech, studies, Harvard, talk, year, congrats, listening, talking, times, paper, dr, interesting

se ejecutó sobre la red de información construida, estableciendo el número de tópicos $K = 50$, previo Dirichlet $\alpha_i=50$ y la distribución de palabras bajo cada tópico $\beta_{ij}=0.01$. Después de estimar la distribución de tópico θ_i para cada usuario (representado por 1,000 tweets agrupados) mediante el algoritmo EM, $\mu_ =30$ se establece como el valor mínimo para la cantidad de usuarios y $\varphi_ =0.65$ como la similitud mínima dentro de la comunidad.

La Tabla 2 muestra la comunidad creada. Dado que no es significativo mostrar los nombres de los usuarios, se muestran únicamente las palabras clave de los tópicos, el nombre de la comunidad, el valor de similitud y el número de cada comunidad.

El resultado de las comunidades se asemeja al interés claro proporcionado por los nodos semilla, los tópicos principales para las comunidades 1 a 4 son: “game and entertainment”, “travel and events”, “Internet and web” y “research and study”.

3.2.3 Clasificación de tweets

Los tweets que se utilizaron para su clasificación fueron de Gemma Petrie (http://twitter.com/GemmaPetrie), quien está interesada en información y medios, eventos sociales y comida. Ella sigue a varias personas que generalmente publican estos temas. Sin embargo, ella también publican tweets que no son de interés para Gemma. Para iniciar el proceso del algoritmo K-Means, se crearon siete tweets de prueba que contenían palabras representativas y relacionadas con estos temas.

Tabla 3

Clasificación de tweets para cada categoría

Categoría	Tweets
Social, Media, News Top ranked words: Social, twitter, media, interesting, people, digital, nytimes, post, information, looking, online, public	Back online after a fantastic weekend with @PeregrinKiwi! Looking fwds to partying again in a month, next time in LA :D Karenwickett, but I wish I could have more face2face conversation with you. Tweetvalue.com calculates your value for twitter, type in username. Mine’s worth $45, much less than the avg $136 for FB If your’e at #asist2010 today, check out my colleague Dave’ talk on social media emergency #KM during the Haiti earthquake Tried to reduce how many people I follow on Twitter and ended up adding 2 more #informationoverloadfail Finished analysis of tweets, status messages, and blogs regarding the type of information provided by user generated content He also said that Zuckerberg rarely posts anything on facebook
Event, Celebration, Activity, Award Top ranked words: Congrats, watching, photo, show, family, nice, afternoon, birthday, life, fun	RT @nmtechcouncil: Reminder: #OpenCoffee this Thursday AM at the Santa Fe Business Incubator –hope to see you there Wow! Congratulations @BAVC for receiving the 2010 MacArthur Award for Creative & Effective Institutions Thanks for sharing and again congrats!	@fstutzman dissertation – networked Information Bejaivor in Life Transition A really funny daily show with Ricky Gervais earlier this week @janedavis @veruka2 Ha! That sounds like quite an evening @clhw1 great weather, lots of family, and a reasonable number of fish Woke up after a great wine with friend... how could I say goobay to all of them and still go to work in the morning?
Food, Drink, Meal, Gourmet Top ranked words: Coffe, food, birthday, beverage, delicious, drink, amazing, favourite, home, beer	Intelligentsia Goes Back to Basic for Brewed Coffe – IF you’re an Intelligentsia regular and drink brewed. @midcenturysal I am unsupervised, eating brisket, drinking drinks named for inventors, and watching bats A Hot Dog for Everyone at The Slaw Dogs – Good Food on the Road on KCRW Ending a long day with warm cookies and cold beer @barbermatt wheeze the juice! Macaroni & Cheese with Blue Cheese, Figs, and Rosemary: Sure to confort the winter blues

Se recolectaron 34,000 tweets de las personas que sigue este usuario; para el resultado de clasificación, cerca de 2000 tweets cayeron dentro de las tres categorías. La Tabla 3 muestra sólo los siete tweets más sobresalientes en cada una de las categorías.

Una vez que el modelo de clasificación está relativamente estable después de varias iteraciones utilizando un conjunto grande de tweets, los indicadores se pueden quedar fijos para que al ingresar nuevos tweets se puedan encontrar rápidamente sus indicadores de interés más cercanos y colocarlos en sus correspondientes categorías.

Conclusiones

Este trabajo propone tareas de investigación empírica, todas ellas sobre el análisis de datos en Twitter: filtrado de tweets basado en los intereses de un usuario, creación de comunidades y clasificación de tweets. Para el filtrado de tweets se recolectan tweets de personas que están siendo seguidas por un usuario y se utilizan para ajustar un modelo LDA como el modelo de interés; después, se utiliza éste para calcular el indicador de interés para los nuevos tweets que se obtienen y de esa forma decidir cuáles se filtran y cuáles se retienen. Para la creación de comunidades, se construye una red de información con usuarios y sus relaciones de seguimiento para poder estimar una distribución de tópicos de interés de usuario utilizando el modelo iTopicModel y con ello realizar una comparación con los indicadores de los tópicos de interés del usuario. Para la clasificación de tweets, se hace una pequeña modificación al algoritmo K-Means que se utiliza para clasificar los tweets en diferentes de categorías. Todas las tareas propuestas están orientadas a mejorar la experiencia del usuario en Twitter para leer sólo los tweets de mayor relevancia para sus intereses personales.

Referencias

[1] Twitter, http://twitter.com

[2] Facebook, http://www.facebook.com

[3] MySpace, http://www.myspace.com

[4] Information Filtering System, http://en.wikipedia.org/wiki/Information_filtering_system

[5] Blei, D.M., Ng, A.Y., & Jordan, M.I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3, 993, 1022.

[6] Yizhou Sun, Jiawei Han, Jing Gao, and Yintao Yu, iTopicModel: Information Network-Integrated Topic Modeling, Proc. 2009 International Conference on Data Mining (ICDM’09), Miami, FL. Dec. 2009.

[7] Bit.ly, http://www.bit.ly

[8] Beta Function, http://en.wikipedia.org/wiki/Beta_Function

[9] K-Means, http://en.wikipedia.org/wiki/K-means_clustering

[10] Wagstaff, K., Cardie C. & Rogers S. (2001). *Constrained k-means clustering with background knowledge*. ICML 2001

[11] Kindermann, R. (2001). *Markov random fields and their applications.*

[12] Cosine Similarity, http://en.wikipedia.org/wiki/Cosine_similarity

[13] Steinbach, M., Karypis, G. Kumar, V. (2000). *A comparison of Document Clustering Techniques*, KDD.

[14] Hoffmann, T. (1999). *Probabilistic latent semantic analysis. In proceedings of UAI*, pp. 289-296.

[15] Wang, R., Jin, R. (2010). An empirical study on the relationships between the followers’ number and influence of microblogging. 2010 International Confenrence.