Redes Sociales y Motores de Búsqueda

Santiago Ricci[1], Esteban Feuerstein[2], Gabriel H. Tolosa[1,2]
sricci.soft@gmail.com; efeurest@dc.uba.ar; tolosoft@unlu.edu.ar

[1]Departamento de Ciencias Básicas, Universidad Nacional de Luján
[2]Departamento de Computación, FCEyN, Universidad de Buenos Aires

Resumen

Internet puede verse desde hace algunos años como un flujo constante de información. Esto se debe en gran parte al contenido producido en las redes sociales. Dicha cuestión puede apreciarse, por ejemplo, en las plataformas de microblogging y dentro de ellas sobre todo en Twitter. Entre las características más importantes de esta plataforma se encuentran los *Trending Topics*, que son un listado de frases, términos y etiquetas relacionados a los temas emergentes más populares en la red social en determinado momento. Así, teniendo en cuenta las cuestiones previamente mencionadas, cabe preguntarse si el hecho de que cierto tema sea *Trending Topic*, influye de algún modo en el volumen de consultas que recibe un motor de búsqueda sobre dicho tema. Este proyecto propone explorar y analizar dicha cuestión y su aplicación en la mejora de los algoritmos de búsqueda.

Palabras clave: redes sociales, motores de búsqueda, trending topics.

Contexto

Esta línea de investigación se encuentra en el marco del proyecto "Modelos y Algoritmos de Búsqueda + Redes Sociales para Aplicaciones Verticales de Recuperación de Información", presentado oportunamente [11]. El mismo pertenece al Departamento de Ciencias Básicas de la Universidad Nacional de Luján.

Introducción

Desde hace ya algunos años, algunos servicios de Internet pueden verse como un flujo de información en tiempo real, lo cual implica agregar una componente temporal a la visión que se tiene de la red [1]. Esto puede apreciarse, por ejemplo, en los servicios de microblogging, los cuales apuntan a una necesidad de comunicación diferente de la de los blogs "tradicionales": la comunicación rápida. Dado que la longitud de las publicaciones se encuentra fuertemente limitada, se requiere menos tiempo para la generación de contenido, lo cual conlleva a que la frecuencia de publicación sea mucho mayor respecto a la de los blogs tradicionales [4]. Un caso concreto de microblogging es Twitter. Este es un servicio basado en lo que se denominan *tweets*. Un tweet es un mensaje cuya longitud máxima es de 140 caracteres. De acuerdo a datos oficiales de Twitter[1], sus usuarios generan más de 95 millones de *tweets* por día, lo cual sugiere la validez de la

[1]http://blog.twitter.com/2010/12/to-trend-or-not-to-trend.html

visión mencionada de Internet.

Los *tweets* poseen una serie de características que son importantes. En primer lugar, pueden contener uno o más símbolos *hashtag* ("#"). Estos son utilizados para etiquetarlos con palabras claves o temas[2]. Por otro lado, en un *tweet* es posible mencionar a otros usuarios. También, es importante destacar a los *Trending Topics* (Tendencias). Esto es un listado de frases, términos y *hashtags* relacionados con los temás más populares en la red social en un determinado momento. En la entrada del blog oficial de Twitter mencionada anteriormente, se afirma que el algoritmo encargado de identificar las tendencias se enfoca principalmente en los temas sobre los que "más se habla" en la inmediatez más que en los temas sobre los que se habló. En otras palabras, intenta identificar las "publicaciones de último momento" y no solamente lo popular; prioriza que el tema sea novedoso sobre la popularidad. Para ello, tiene en cuenta dos factores: el volumen de los términos mencionados en Twitter y su velocidad en el crecimiento. Por defecto, las tendencias se determinan de forma personalizada para cada usuario en base a los usuarios que sigue y a su localización. Sin embargo, un usuario puede optar por obtener *Trendings Topics* de forma no personalizada, en base a su región geográfica.

Sobre el escenario de Internet también se encuentran los motores de búsqueda, herramientas indispensables para el acceso al contenido en la red. Teniendo en cuenta además, que los datos e información en la web son cada vez más ricos y complejos, se hace necesario disponer de los mismos en tiempo y forma. En consecuencia, la búsqueda se ha convertido en un proceso clave y central en la web. Esto se traduce en que los motores de búsqueda deban responder millones de consultas (*queries*) por día, lo cual implica eficiencia y efectividad para poder otorgar a los usuarios respuestas relevantes lo más rápido posible [12].

Existen diversos estudios acerca de Twitter. En [4] se estudia al servicio desde el punto de vista estructural y del contenido. Una publicación posterior [6] amplía dicha caracterización. En [7] se exponen las características topológicas del servicio de microblogging y se sugiere que gran parte (85 %) de los *Trending Topics* estan relacionados con las noticias del momento o con aquellas que son persistentes en el tiempo y que una gran porción (31 %) de los mismos dura aproximadamente un día. Además se concluye que los usuarios tienden a "hablar" sobre dichas noticias. En [2], también estudian los *Trending Topics* y se afirma que aquellos con grandes duraciones están caracterizados por la naturaleza "resonante" del contenido de sus *tweets* asociados, el cual proviene, generalmente, de los medios de comunicación tradicionales. De este modo, Twitter se comporta como un amplificador selectivo del contenido generado por los medios tradicionales mediante cadenas de *retweets*. En [10] se compara la tarea de búsqueda de los usuarios en Twitter respecto a la misma en los motores de búsqueda y se concluye que los usuarios utilizan Twitter para monitorear un tema y la Web para aprender más acerca del mismo.

También, es conocido el uso de esta red para expresar opiniones acerca de diferentes temas, lo cual se ha traducido en gran cantidad de artículos que plantean diferentes enfoques sobre cómo realizar Minería de Opinión sobre la red social, como por ejemplo, [14] o [9].

Sin embargo, aún no está clara la relación entre los motores de búsqueda y las redes sociales, por lo que en esta línea de investigación se trabaja sobre las siguientes cuestiones:

1. El hecho de que cierto tema sea *Trending Topic*, ¿Influye de algún modo en el volumen de consultas que recibe un motor de

[2]https://support.twitter.com/articles/49309-what-are-hashtags-symbols#

búsqueda sobre dicho tema?

2. ¿Es posible emplear esta información de la red social para "optimizar" el algoritmo de búsqueda, por ejemplo, utilizándolo para refinar las políticas de reemplazo de una memoria caché o el *prefetching* [5] de resultados?

3. De igual manera, ¿Se podría aplicar dicha información en el proceso de desambiguación del sentido de una consulta?

Líneas de investigación y desarrollo

Las hipótesis presentadas en la sección anterior sugieren las siguientes líneas de investigación:

1. **Caracterización de los Trending Topics:** en esta línea se pretende determinar propiedades de los *Trending Topics* que son útiles en relación con las temáticas propuestas. Una primer problemática es clasificar el contenido que intentan describir los *Trending Topics*, dado que es conocido que no todos los usuarios utilizan la red social con los mismos objetivos [4] y el espectro de temas abarcados por la misma es virtualmente ilimitado. En este sentido, también es importante estudiar los períodos en los que se encuentran activos, sus repeticiones en el tiempo, el porcentaje de usuarios que participan en los mismos, entre otras cuestiones. Aquí también puede enmarcarse el proceso de derivación de consultas desde los *Trending Topics*. Este proceso puede ser determinante en los resultados, dado que no siempre está claro el tema que representa. El *parsing* del mismo puede ser problemático e incluso generar consultas ambiguas.

2. **Caracterización del volumen de tráfico en un motor de búsqueda mediante métodos indirectos:** esta línea pretende intentar establecer el impacto en la carga de trabajo de un conjunto de consultas a un motor de búsqueda. Como esta información es propietaria de los proveedores de los servicios de búsqueda (se los caracteriza como "ambientes no cooperativos" [8]), se requieren métodos "externos" que permitan obtener aproximaciones válidas. Este problema ha sido abordado en el área de Recuperación de Información Distribuida para obtener descripciones de los recursos objetivo (por ejemplo, Query Based Sampling [3]). Esta idea ya ha sido explorada inicialmente en nuestro proyecto (se presenta en la próxima sección).

3. **Algoritmos que exploten los *features* de los *Trending Topics* para la desambiguación del sentido de una consulta:** múltiples elementos se tienen en cuenta a la hora de decidir el ranking final de una consulta para un usuario particular. Se sabe que su historial de búsqueda e información de perfil son indicadores de preferencias temáticas, utilizados por ejemplo, para el proceso de desambiguación. Se considera explorar si la información proveniente de una red social aporta a este proceso.

4. **Modelos de *caching* y *prefetching* que usan la información de los Trending Topics para tomar decisiones que mejoren la performance:** estas dos técnicas son ampliamente usadas en el ámbito de los motores de búsqueda y permanentemente se estudian nuevas estrategias. El uso propuesto puede permitir mejoras particulares bajo ciertas circunstancias, por ejemplo, búsquedas en tiempo real.

Resultados y objetivos

Sobre estas líneas de investigación se ha comenzado a trabajar y se han obtenido algunos resultados prometedores. Se realizó una experiencia de pequeña escala sobre los *Trending Topics* de Twitter para Argentina durante una semana. Por otro lado, se observaron las tendencias de evolución de algunas consultas mediante Google Trends y se aplicaron tres métricas que apuntan a cuantificar el impacto de la red social en un motor de búsqueda:

1. **Variación porcentual del interés de la consulta derivada para un *Trending Topic* respecto al día anterior:** esta métrica intenta capturar el hecho de que si los *Trending Topics* influyen en el volumen de consultas, entonces debe existir una diferencia significativa en el interés en dicha consulta cuando el tema es *Trending Topic* y cuando no lo es.

2. **Cuantifiación del cambio en la tendencia que experimenta cierta consulta derivada cuando el tema se convierte en *Trending Topic*, respecto a su tendencia en los n días anteriores.**

3. **Algoritmo de detección de picos *(burst detection)*:** este algoritmo fue utilizado de acuerdo a [13]. Aquí se intenta demostrar que el día que un tema es *Trending Topic*, registra un pico en la aparición de consultas en el motor de búsqueda.

Los resultados preliminares muestran que aproximadamente el 70 % de las consultas aumentan su volumen de tráfico cuando aparecen como *Trending Topic* en Twitter y en aproximadamente el 60 % de los casos este crecimiento supera el 40 % respecto al día anterior.

Formación de Recursos Humanos

Esta línea de investigación surgió de una Pasantía Interna Rentada (PIR) de la UNLu que el primer autor realizó en el proyecto en que está inserta (Ver "Contexto") bajo la supervisión de Prof. Gabriel Tolosa. Durante dicha pasantía surgieron varias direcciones futuras, por lo que el primer autor va a desarrollar su Trabajo Final de Licenciatura en Sistemas de Información en esta área y se prevé incorporar un nuevo pasante durante el año en curso.

Referencias

[1] Agarwal, S.; Agarwal, S. Social networks as Internet barometers for optimizing content delivery networks. In 3rd International Symposium on Advanced Networks and Telecommunication Systems (ANTS). 2009.

[2] Asur, S.; Huberman, B. A.; Szabo, G.; Wang, C. Trends in social media: Persistence and decay. In 5th International AAAI Conference on Weblogs and Social Media. 2011.

[3] Callan, J.; Connel, M. Query-based sampling of text databases. In ACM Transactions on Information Systems, v. 19, n. 2, pp. 97-130. 2001.

[4] Java, A.; Song, X.; Finin, T.; Tseng B. Why we twitter: understanding microblogging usage and communities. In Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis (WebKDD/SNA-KDD '07), pp. 56-65. 2007

[5] Jonassen, S.; Barla Cambazoglu B.; Silvestri F. Prefetching query results and its impact on search engines. In Proceedings of

the 35th international ACM SIGIR conference on Research and development in information retrieval (SIGIR '12), pp. 631-640. 2012.

[6] Krishnamurthy, B.; Gill, P.; Arlitt M. A few chirps about twitter. In Proceedings of the first workshop on Online social networks (WOSN '08), pp. 19-24. 2008.

[7] Kwak, H.; Lee, C.; Park, H.; Moon, S. What is Twitter, a social network or a news media?. In Proceedings of the 19th international conference on World wide web (WWW '10), pp. 591-600. 2010.

[8] Luo S. Federated search of text search engines in uncooperative environments. PhD Thesis. Carnegie Mellon University. 2006.

[9] Meng X.; Wei, F.; Liu, X.; Zhou, M.; Li, S.; Wang, H. Entity-centric topic-oriented opinion summarization in twitter. In Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '12), pp. 379-387. 2012.

[10] Teevan, J.; Ramage, D.; Ringel Morris, M. #TwitterSearch: a comparison of microblog search and web search. In Proceedings of the fourth ACM international conference on Web search and data mining (WSDM '11), pp. 35-44. 2011.

[11] Tolosa, G. H.; Bordignon F. Modelos y algoritmos de búsqueda + redes sociales para aplicaciones verticales de recuperación de información. En Proceedings del XIII Workshop de Investigadores en Ciencias de la Computación (WICC 2011), pp. 243-247. 978-950-673-892-1. 2011.

[12] Tolosa G. H.; Feuerstein E. Mejoras algorítmicas y estructuras de datos para búsquedas altamente eficientes. En Proceedings del XIV Workshop de Investigadores en Ciencias de la Computación (WICC 2012), p. 740-744. 978-950-766-082-5. 2012.

[13] Vlachos, M.; Meek, C.; Vagena, Z.; Gunopulos, D. Identifying similarities, periodicities and bursts for online search queries. In Proceedings of the 2004 ACM SIGMOD international conference on Management of data (SIGMOD '04), pp. 131-142. 2004.

[14] Wang, X.; Wei, F.; Liu, X.; Zhou, M.; Zhang, M. Topic sentiment analysis in twitter: a graph-based hashtag sentiment classification approach. In Proceedings of the 20th ACM international conference on Information and knowledge management (CIKM '11) pp. 1031-1040. 2011.