
Propuesta de estudio del campo semántico de los libros electrónicos en Twitter

The 'electronic book' semantic field in Twitter

José Luis Alonso Berrocal (1), Raquel Gómez Díaz (2), Carlos G. Figuerola (1), Ángel F. Zazo Rodríguez (1), Jose Antonio Cordón García (2)

(1) Departamento de informática y automática, Universidad de Salamanca. (2) Departamento de biblioteconomía y documentación, Universidad de Salamanca. {berrocal|rgomez|figue|zazo|jacordon}@usal.es

Resumen

Las redes sociales han conseguido que la Red sea un repositorio de información de la que se pueden extraer múltiples informaciones. Twitter es una de las más conocidas y de las que mayor incremento en número de usuarios ha tenido en los últimos años. El análisis del contenido de sus mensajes proporciona valiosa información sobre los autores de los tweets, las relaciones entre seguidores y seguidos, etc, así como aquellos mensajes que marcan tendencia. En el caso concreto de este trabajo esta información se aplica al campo semántico de los libros electrónicos.

Palabras clave: Twitter. Análisis semántico. Libros electrónicos.

Abstract

Social networks have transformed the Web into a repository of information from which very diverse information can be extracted. Twitter is one of the best known and has had one the greatest increase in number of users of the recent years. Content analysis of their messages provides valuable information about the authors of the tweets, the relationship between followers and the followed ones, etc. as well as about messages that set the trend. In the case of this paper, this kind of research is applied to the semantic field of electronic books.

Keywords: Twitter. Semantic analysis. Electronic books.

1. Introducción

Las redes sociales, los blog, o cualquier foro online han convertido a Internet en un gran repositorio de comentarios de los que se pueden extraer múltiples informaciones (Thelwall, 2011) tal y como lo muestran los numerosos trabajos de investigación que se vienen realizando en los últimos años.

De todas las redes sociales nos interesa el caso de Twitter, que es una de las más conocidas y de las que mayor incremento en número de usuarios ha tenido en los últimos años (The Cocktail Analysis, 2011) consolidándose como una herramienta fundamental en las estrategias de comunicación y difusión de la información.

La característica principal de este modo de comunicación reside en la longitud de sus mensajes -de 140 caracteres como máximo-, llamados tweets, que se muestran en la página principal del usuario. Los usuarios pueden suscribirse a los tweets de otros usuarios, lo que se denomina "siguiendo" y a los suscriptores se les llaman "seguidores". Por defecto, los mensajes son públicos, pudiendo difundirse de manera privada mostrándolos únicamente a seguidores. Este método se utiliza para compartir información y también para describir prácticamente cualquier actividad diaria (Java, 2009), ya que permite

que los usuarios puedan expresar sus opiniones e intereses, de forma abreviada y altamente personalizada en tiempo real (García Esparza, 2012). Su importancia se muestra en que está presente en prácticamente todas las áreas de la vida (social, económico, educativo...) y abarca cualquier tema (deportes, cultura, ocio, industria, ciencia).

Si valoramos la importancia de esta red social en términos cuantitativos, es necesario hacer referencia al volumen de tweets que se generan a diario -en junio de 2011 unos 200 millones-, cifra que va en aumento- (Lee, 2011). Si la valoración la hacemos en términos cualitativos es necesario considerar cómo su influencia se refleja en la multitud de actos sociales que se retrasmiten en tiempo real con el objetivo de ganar en visibilidad así como en el gran número de mensajes no originales que se vuelven a difundir (retweets), por lo que se puede considerar incluso que Twitter llega a constituir nichos de opinión, ya que un mensaje creado por una persona (ya sea original o un fragmento de otra obra como un titular de un periódico o un extracto de una noticia) puede ser retwitteado por otro u otros usuarios que a su vez lo vuelven a retransmitir provocando un efecto de difusión en forma de racimo.

Es cierto que gran parte de la información que aportan los tweets es completamente irrelevante, además es necesario considerar que en muchos casos los mensajes aislados de su contexto pierden valor, pero también es una fuente de información muy rica porque recoge de manera condensada la información relevante para los usuarios, ya sean personas individuales, instituciones o empresas, que son los que destacan noticias, opiniones o sentimientos, información que resultaría muy difícil de recopilar por otros canales y precisamente por esto, el análisis de los mensajes de twitter está siendo utilizado como materia de base para múltiples investigaciones que van desde el papel que juegan los distintos tipos de usuarios en la difusión de la información (Cha, 2012) a análisis sociológicos (Chen, 2012) tendencias en el aprendizaje (The Semantic Web, 2012), aplicaciones a la clasificación (Lee, 2011) y recuperación de la información (García Esparza, 2012 ; Yerva, 2012), análisis semántico (Narr, 2011) o aplicados a sectores concretos como el caso de la edición (Uso, 2011).

2. Objetivos

En este trabajo pretendemos realizar una propuesta de estudio de campos semánticos en twitter. Dado un campo semántico y una vez definidos unos parámetros de análisis (simetría, influencia de los perfiles según los seguidores, estudio de entidades, retweets, análisis léxico y semántico del contenido y teoría de grafos) será posible abordar el estudio de cualquier campo semántico presente en twitter.

Se trata de una propuesta de trabajo inicial, que tendrá que tener una continuidad en el tiempo para poder ir perfilando, corrigiendo y ampliando dicha propuesta hasta obtener un modelo que nos permita realizar los estudios con garantías.

Con el objeto de ejemplificar los parámetros de estudio se ha seleccionado el campo semántico "libros electrónicos", un tema de especial relevancia por sus implicaciones sociales, educativas, tecnológicas, económicas (Cordón, 2011) pero suficientemente concreto como para abordar un estudio de estas características.

3. Metodología

El trabajo con twitter se ha llevado a cabo analizando la API que suministra la red social. La API de twitter ofrece tres posibilidades:

- *Streaming API*: Proporciona un subconjunto de tweets casi en tiempo real. Se establece una conexión permanente por parte del usuario y se recibe un flujo continuo de

tweets en formato json. Se puede obtener una muestra aleatoria (*statuses/sample*), un filtrado (*statuses/filter*) por palabras claves o por usuarios. Sin embargo, los métodos más interesantes cómo obtener todo el caudal de tweets (*statuses/firehose*) o sólo los tweets que tienen enlaces (*statuses/links*) o los tweets con retweets (*statuses/retweet*) no siempre están disponibles. El flujo es continuo y la velocidad de recepción de tweets tiene fluctuaciones que dependerán del ancho de banda de los dos extremos de la conexión y la sobrecarga de los servidores de Twitter.

- *Search API*: Suministra los tweets con una profundidad en el tiempo de 7 días que se ajustan a la consulta solicitada. Es posible filtrar por cliente utilizado, lenguaje y localización. No requiere autenticación y los tweets se obtienen en formato json o atom. Existe una limitación de 150 peticiones a la hora por usuario o por IP si la llamada no está autenticada y si se ajusta adecuadamente el programa podemos obtener un máximo de 1500 tweets.
- *REST API*: Ofrece a los desarrolladores el acceso al core de los datos de Twitter. Todas las operaciones que se pueden hacer vía web son posibles realizarlas desde el API. Dependiendo de la operación requiere o no autenticación, con el mismo criterio que en el acceso web. Soporta los formatos: xml, json, rss, atom. Podemos obtener un máximo de 3200 tweets y se pueden realizar peticiones acotadas en el tiempo, así como obtener los following (seguidos) y los followers (seguidores) de un usuario dado.

Este trabajo se basa en REST API y para ello se han desarrollado los programas que nos permiten el trabajo con la API y se han almacenado los datos en Redis y en CouchDB para su posterior tratamiento y obtención de los datos.

Para el trabajo con la API es necesario tener una cuenta de desarrollador, que se abre desde la dirección <https://dev.twitter.com/>. Una vez obtenida la cuenta correspondiente se trabajará con el lenguaje de programación que se desee. En este sentido conviene considerar que existen librerías para trabajar con la API para numerosos lenguajes de programación (<https://dev.twitter.com/docs/twitter-libraries>).

La mayor parte de la información que se obtiene de twitter está en el formato denominado JSON, acrónimo de JavaScript Object Notation, que es un formato ligero para el intercambio de datos. Se emplea habitualmente en entornos donde el

tamaño del flujo de datos entre cliente y servidor es de vital importancia (Introducción, 2012).

Este hecho ha determinado la elección de las dos herramientas de almacenamiento seleccionadas. Por un lado, REDIS (<http://redis.io>), una base de datos NoSQL, de tipo clave-valor. REDIS da soporte para listas, conjuntos, conjuntos ordenados y hashes, absolutamente esenciales en el tratamiento de la información de twitter. El rendimiento es muy alto y la flexibilidad y la adaptación al formato JSON excepcional.

Por otro lado, se ha utilizado CouchDB (<http://couchdb.apache.org/>), una base de datos orientada a documentos, sin estructura fija, lo cual permite guardar de manera conjunta documentos con distintos campos dentro de la misma base de datos. Cada documento se almacena en formato JSON, adaptándose perfectamente al formato de twitter. Una funcionalidad interesante es que permite la sincronización de sus datos de forma muy sencilla. De este modo, es posible obtener fácilmente una o más réplicas de nuestra base de datos para implementar arquitecturas de alta disponibilidad o de balanceo de carga.

Con toda esta infraestructura se procedió a recoger la información de un conjunto de perfiles de usuario, definidos por especialistas en la materia, en este caso de libros electrónicos, y que han servido como base de estudio de la propuesta presentada.

La información recogida pertenece a las siguientes categorías: seguidos y seguidores, los tweets y los datos de los perfiles de las siguientes cuentas:

24symbols	actualidad_ed
anatomiaedicion	editordigital
electragir	futuroslibro
hubertguillaud	jacordongarcia
jafurtado	javierelaya
lecturalab	libroscom
librosybitios	pubperspectives
territorioebook	

Tabla 1. Perfiles de usuario del estudio

Con todos los datos se procedió al análisis de diferentes elementos que servirán para caracterizar el campo semántico objeto de estudio.

3.1. Simetría (amigos)

La simetría permite medir el número de amigos mutuos que tiene un determinado perfil. Para ello se calcula el número de seguidos y el número de seguidores. Se calcula qué seguidos de los obtenidos no siguen al perfil de estudio y los seguidores que no son amigos del perfil estudiado. Finalmente se hace una intersección de ambos conjuntos y obtenemos el número de amigos mutuos.

Para perfiles de tipo profesional es un modelo perfectamente válido pues el tamaño de las muestras se puede considerar como altamente significativo del comportamiento global del perfil.

Si se desea trabajar con perfiles que tienen un tamaño de población muy elevado el tamaño de muestra puede presentar desviaciones significativas estadísticamente.

Conocer los seguidos que no siguen a un determinado perfil puede representar una información interesante en el establecimiento de los flujos de información entre los perfiles.

Relacionado con ello otro parámetro interesante es obtener la similitud de los perfiles en función de los seguidos y de los seguidores comunes que se tienen. Requiere la comprobación de todos los perfiles con todos los perfiles, lo que requiere crear una matriz de $N \times N$ elementos para calcular los niveles de similitud.

3.2. Influencia de los perfiles según los seguidores

Un aspecto interesante a calcular en Twitter es la influencia que pueden tener los perfiles. Una medida de gran importancia para medir la influencia es calcular los retweets de un tweet. La influencia será directamente proporcional al número de retweets que se hagan. Este aspecto se explicará posteriormente, pero antes se puede calcular de forma más simple una medida de influencia según el número de seguidores de los seguidores de un perfil. Esto puede dar la influencia potencial, aunque requiere que se realice un recolector para obtener dicha información.

Es posible recoger N niveles de seguidores para poder calcular la influencia de los mismos. En el caso de este estudio se trabaja con un nivel de seguidores, que nos permita calcular la media de seguidores de los seguidores de un perfil.

3.3. Estudio de entidades

Las entidades agrupan 3 tipos esenciales de información que aparece en los tweets.

- *Entidades de usuario*: Comienzan con el símbolo @ y representan reply, respuestas al mensaje de otros usuario, o retweets, difusión de mensajes interesantes.
- *Hashtag*: Son palabras clave para poder etiquetar los tweets. Comienzan por el carácter # que va seguido de la etiqueta (#etiqueta). Se utilizan para poder agrupar información relacionada y sobre todo en los eventos en los que se utiliza Twitter.
- *URL*: Se trata de direcciones web añadidas a los tweets, normalmente en formato abreviado para ahorrar caracteres.

Aparte de los valores obtenidos para cada una de las entidades, que caracterizan más los tweets analizados se puede completar la información comprobando de las entidades de usuario existentes cuáles son los perfiles que estamos siguiendo, lo que determina las cohesiones existentes en las entidades de usuario analizadas. Este sistema permite realimentar el proceso de recogida del campo semántico que se pretenda analizar.

3.4. Retweets

La retransmisión de tweets por parte de un perfil, es decir, la difusión de la información considerada más interesante para ese perfil, permite diferentes estudios, entre ellos determinar los aspectos más interesantes que la comunidad considera dignos de ser difundidos. En esta fase del trabajo se analizan el número de retweets que se realizan a los seguidos y si existen autocitas en los retweets.

Para analizar la influencia de los tweets de un determinado perfil, se calculan los retweets que se han realizado, poniendo en relación, la ratio tweet/retweet y el número total de retweets que ha tenido frente al número de tweets que tiene el perfil.

Esto determinará la importancia de los mensajes de los perfiles en función de la importancia que se le da por parte de otros perfiles.

3.5. Análisis léxico

Otro aspecto que se puede tener en cuenta es el análisis léxico y semántico de los tweets. Aquí se puede desplegar toda la batería de estudios tanto desde el punto de vista terminológico como desde el ámbito de la recuperación de información.

En este trabajo se aborda solamente un estudio de la nube de etiquetas que se obtiene, y que permite caracterizar los tweets estudiados.

3.6. Teoría de grafos

Cuando se trabaja con redes (twitter es una red), se utiliza una rama de las matemáticas llamada Teoría de grafos para la definición de los conceptos. Aquí simplemente se definen algunos de los conceptos que se pueden utilizar en un estudio como el que aquí se desarrolla.

En primer lugar es necesario tener en cuenta que se entiende por grafo un conjunto de vértices y un conjunto de líneas entre pares de esos vértices.

Este grafo permite representar adecuadamente la estructura de una red, donde los vértices o elementos de la red se denominan genéricamente nodos (siendo la unidad más pequeña de la red) y estos nodos se encuentran comunicados mediante líneas. Estas líneas pueden ser dirigidas, el sentido de la conexión es importante, denominándose arcos; o bien líneas no dirigidas, la conexión indica un sentido bidireccional, que se denominan aristas.

De esta forma, se pueden tener grafos dirigidos o digrafos cuando contienen arcos y grafos no dirigidos cuando contienen aristas.

Los grafos dirigidos finitos con n nodos, se representan como estructuras de datos por medio de una matriz de adyacencia: una matriz n -por- n cuyas entradas en la fila i y la columna j dan el número de arcos desde el nodo i -ésimo al j -ésimo. Como se muestra en el siguiente ejemplo (Figura 1, Tabla II).

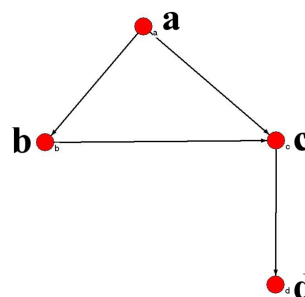


Figura 1. Grafo de ejemplo

a_{ij}	a	b	c	d
a	0	1	1	0
b	0	0	1	0
c	0	0	0	1
d	0	0	0	0

Tabla II. Matriz de adyacencia del grafo

Formalmente se define la matriz de adyacencia como $V = \{v_1, v_2, \dots, v_n\}$ de forma que, como se

puede apreciar, aparece un 1 cuando existe el enlace y un 0 en caso contrario.

A partir de esa representación en matriz se pueden aplicar infinidad de cálculos, muchos de los cuales generan índices y algunos de ellos son los que van a ser empleados en este estudio.

Las estructuras de enlaces, una vez transformadas en grafos y matrices de adyacencia, permiten discernir los patrones estructurales del sitio. De esta forma, las estructuras hipertextuales de una red serán diferentes dependiendo de su funcionalidad dentro de la propia Red. Según sea la función de la red éste tendrá una estructura de enlaces determinada que compartirá con otros sitios similares.

Asimismo se pueden obtener índices o medidas que afectan a toda la red, a los nodos individualmente, destacando las denominadas medidas de centralidad, o las medidas de posicionamiento, entre las que se encuentra el Page-Rank.

En el caso concreto de este trabajo simplemente se utilizarán las conocidas como medidas de centralidad.

3.6.1. Centralidad

Las medidas de centralidad permiten estudiar qué nodos son los más centrales, los más importantes, los que poseen un mayor poder o bien los más prestigiosos. Las medidas de centralidad son una buena aproximación al análisis de los grafos, y permiten evaluar las dimensiones reales del prestigio y del poder. Según Freeman (1979) la clasificación es la siguiente:

- El grado o degree es el número de enlaces directos que un nodo tiene. En los grafos no dirigidos este es un valor único, en el que se tienen en cuenta que los enlaces son de doble sentido, pero en el caso de los grafos dirigidos, con los que estamos trabajando podemos tener dos variantes:
 - *Grado de entrada*, indegree, indica el número de nodos que se relacionan (contactan) con un nodo determinado. Esto permite conocer los nodos más prestigiosos o de referencia para el resto de la red. Determina el prestigio y poder a nivel local. Cuando este valor se normaliza por el número de nodos total este concepto se redefine como visibilidad.
 - *Grado de salida*, outdegree, indica el número de relaciones, de enlaces, directos iniciados desde un nodo determinado. Se refleja tanto la actividad social, como su

capacidad para poder acceder al resto de los nodos. Cuando este valor se da normalizado se redefine como luminosidad.

- *Grado de intermediación*: El grado de intermediación, betweenness, nos indica hasta qué punto un nodo determinado está en una posición intermedia entre el resto de los nodos. En el cálculo de esta medida intervienen las distancias geodésicas (más cortas) entre los diferentes nodos. Los nodos con un mayor índice tienen un gran poder, al controlar los flujos de información más óptimos.
- *Grado de cercanía*: El grado de cercanía, closeness, mide la distancia media de cada nodo con el resto de nodos. Los resultados más altos sugieren una mayor facilidad de acceso al resto de los elementos que conforman el grafo y por lo tanto tienen una mayor capacidad para obtener y enviar información.

Se pueden tener nodos que se encuentren poco conectados con el resto (y por lo tanto un índice de intermediación bajo), pero que si se conectan a un nodo importante pueden tener un alto grado de cercanía y por lo tanto, si llevamos este concepto al terreno político, se trataría de elementos con una gran influencia.

Conviene señalar que este estudio del grafo es transversal y se puede añadir de forma complementaria al resto de apartados estudiados.

4. Resultados

Una vez analizados cada uno de los parámetros se muestra su aplicación al caso concreto de los libros electrónicos

4.1. Número de tweets

Esta medida define el nivel de actividad de la cuenta. Se han recuperado los últimos tweets que nos suministra la API, por ello el tope es de 3200. En la Tabla III se muestran los resultados obtenidos. Lo más significativo son los valores bajos, los perfiles con menor actividad puesto que para los valores altos el tope no permite conocerlo en detalle. Para próximas recogidas hay que obtener el número de tweets real para poder determinar correctamente el nivel de actividad de los perfiles.

4.2. Número de seguidos y seguidores

El número de seguidos y seguidores (Tabla III) nos permite determinar el nivel de relaciones de los perfiles de usuario.

Perfil	Tweets	Seguidos	Seguidores
24symbols	627	362	4157
actualidad_ed	3200	373	7945
anatomiaedicion	1455	1376	3377
editordigital	374	574	2365
electragir	47	40	27
futuroslibro	261	43	2974
hubertguillaud	2033	297	3667
jacordongarcia	798	205	286
jafurtado	3199	3372	10399
javiercelaya	3199	412	11393
lecturalab	3200	87	1906
libroscm	1750	2028	1965
librosybitios	2287	126	2661
pubperspectives	3200	5961	14716
territorioebook	2888	162	1259

Tabla III. Número de tweets, perfiles seguidos y seguidores

Las cuentas con mayor número de seguidos son pubperspectives, jafurtado, libroscm y anatomiaedicion.

Sobre los seguidores el orden es pubperspectives, javiercelaya, jafurtado y actualidad_ed.

Junto a estos datos se obtuvo el grafo completo de relaciones existentes incorporando al grafo todos los seguidos para cada uno de los perfiles. Una vez obtenido el grafo de relaciones pudimos calcular los diferentes parámetros de grafo. De los datos obtenidos el más significativo es el índice de intermediación (Figura 2) que da como resultado que el perfil que tiene el valor más alto es pubperspectives, seguido a distancia por libroscm.

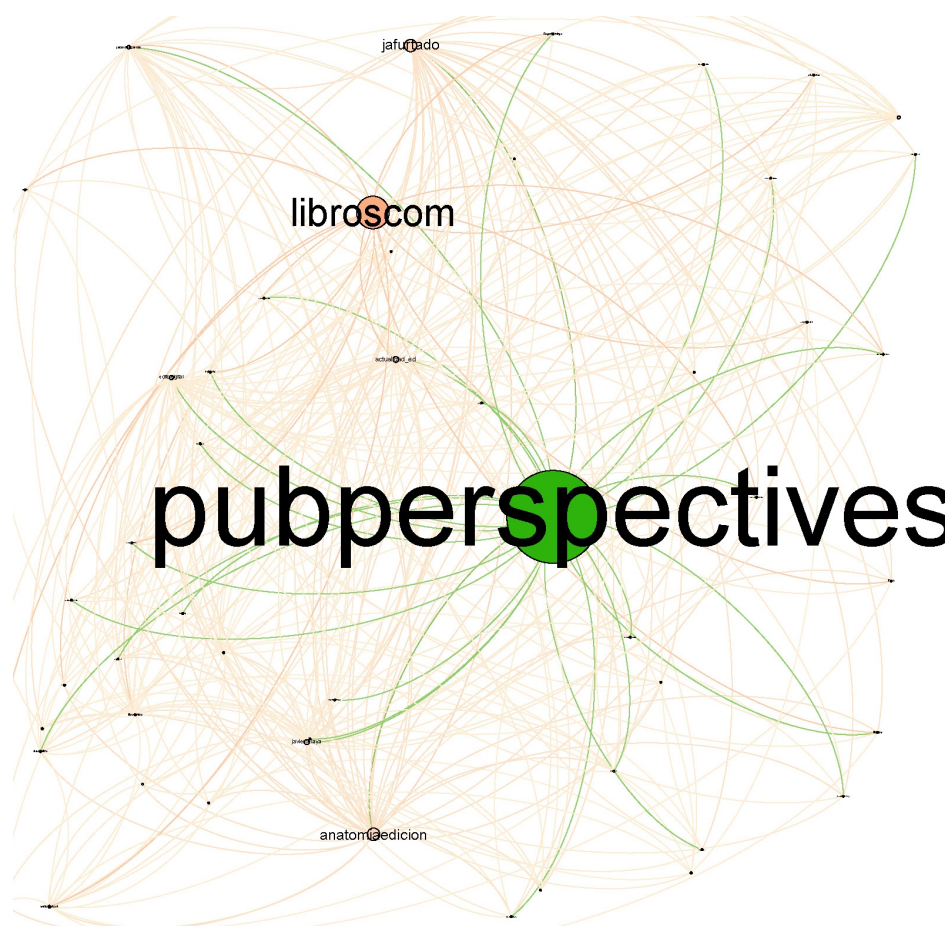


Figura 2. Representación del grado de intermedación.

En las Figuras 3 y 4 se muestran las relaciones directas más importantes de ambos perfiles.

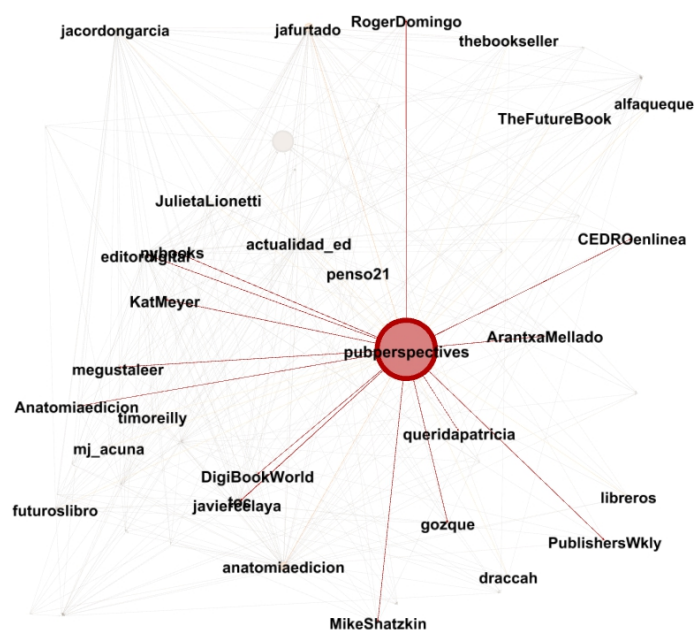


Figura 3. Conexiones del perfil pubperspectives

En el trabajo de futuro se estudiará este sistema de realimentación automático.

4.3. Simetría (amigos)

En la Tabla IV representamos el número de amigos que tiene cada uno de los perfiles. Aunque se tiene el número real de amigos, resulta más importante el porcentaje de amigos en relación con el de los seguidos. El dato más significativo es el del perfil jafurtado con un 99% de amigos, que define un grupo muy compacto de relaciones. Así mismo pubperspectives, a pesar de ser eje del flujo de relaciones tiene un menor porcentaje de amigos.

24symbols	163	45,03%
actualidad_ed	253	67,83%
anomiaedicion	778	56,54%
editordigital	373	64,98%
electragir	8	20,00%
futuroslibro	26	60,47%
hubertguillaud	170	57,24%
jacordongarcia	159	77,56%
jafurtado	3336	98,93%
javiercelaya	186	45,15%
lecturalab	50	57,47%
libroscom	1159	57,15%
librosybitios	80	63,49%
pubperspectives	2102	35,26%
territorioebook	67	41,36%

Tabla IV. Número de amigos y su porcentaje sobre el total

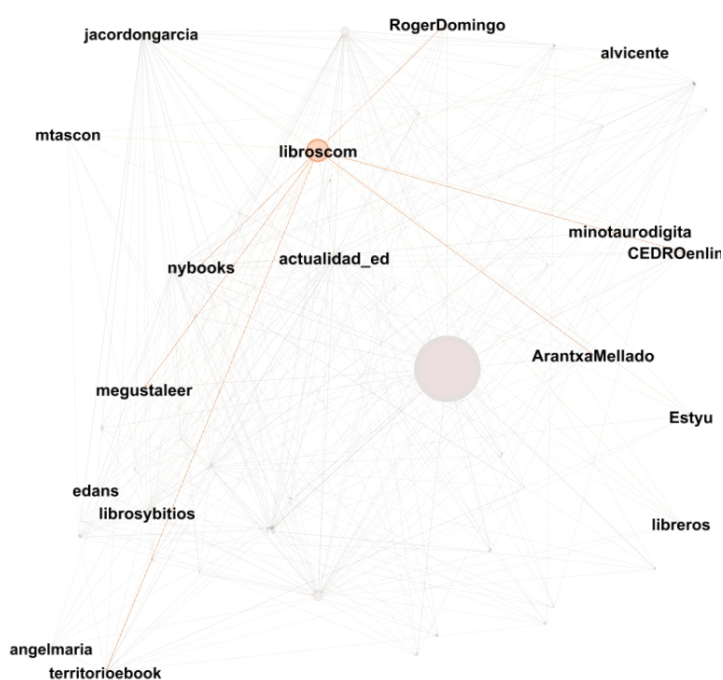


Figura 4. Conexiones del perfil libroscom

Los datos obtenidos pueden servir para realimentar el mecanismo de obtención de datos, ampliando así el número de perfiles que servirán para obtener la muestra, pero además con perfiles directamente relacionados con los que mayor grado de intermediación tienen y siendo, por lo tanto, de calidad suficiente.

4.4. Influencia de los seguidores

En la Figura 5 representamos la influencia de los seguidores tanto para el total de seguidores del perfil como para el top10 de los mismos.

Es muy destacable que el perfil jafurtado se encuentra en posición máxima de influencia teniendo en cuenta cualquiera de ambas medidas.

Es más precisa y adecuada la media total, porque como podemos ver en el top10 aparece el perfil electragir como máximo exponente, cuando en estos momentos su influencia no es en absoluto de esa magnitud.

24symbols	50 %
actualidad_ed	86 %
anatomiaedicion	81 %
editordigital	60 %
electragir	0 %
futuroslibro	89 %
hubertguillaud	61 %
jacordongarcia	69 %
jafurtado	92 %
javiercelaya	78 %
lecturalab	46 %
libroscm	79 %
librosybitios	81 %
pubperspectives	69 %
territorioebook	70 %

Tabla V. Entidades de usuario seguidos

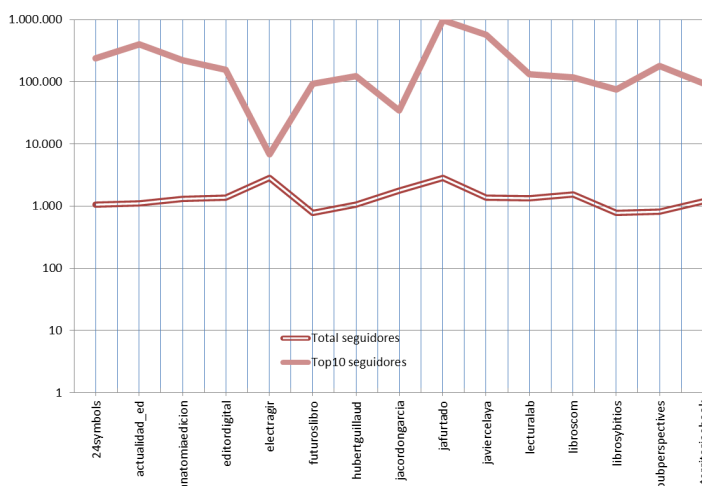


Figura 5. Influencia de los seguidores

4.5. Entidades

En cuanto a la proporción de las entidades utilizadas en los tweets, podemos concluir que el 65% son del tipo @, el 31% de tipo # y 4% se trata de enlaces.

Más interesante es el grafo obtenido con las entidades enlazadas desde las cuentas con las que hemos trabajado (Figura 6).

Podemos realizar un procesado completo de dicho grafo y obtener algunos datos interesantes de dicho procesamiento.

Las entidades más enlazadas son territorioebook, actualidad_ed, librosybitios y libroscm.

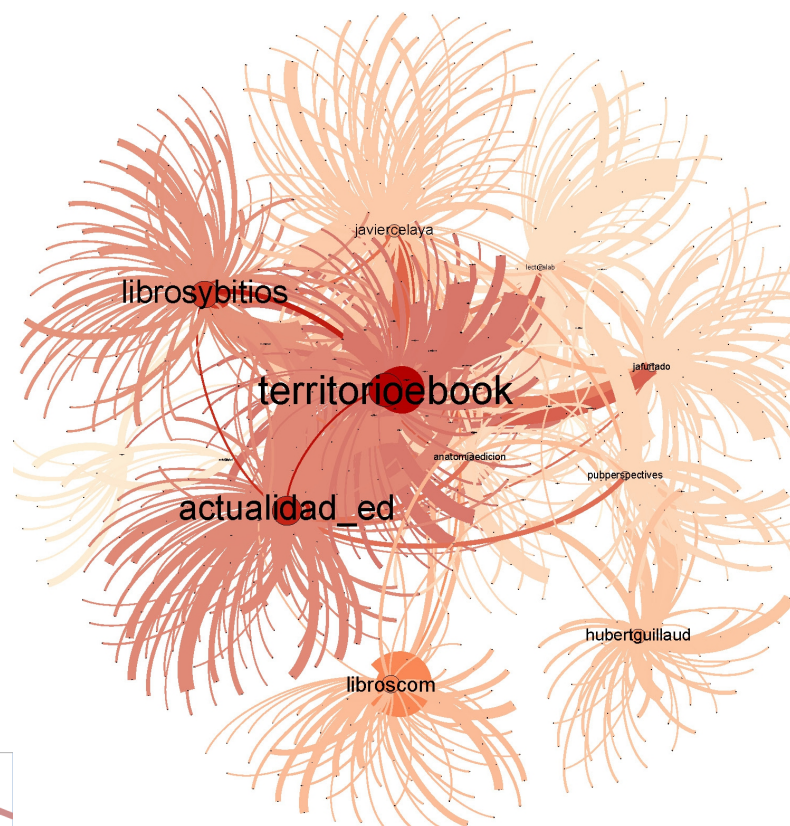


Figura 6. Representación del grado de entrada de las entidades @

El perfil que más entidades tiene en sus tweets es javiercelaya, de forma significativa con respecto al resto.

Analizando el grado de intermediación del grafo obtenido destaca claramente javiercelaya como eje del flujo de información respecto a las entidades, seguido a cierta distancia por actualidad_ed y jafurtado.

4.5.1. Entidades de usuario seguidos

Se puede completar este estudio de las entidades (Tabla V), indicando de las entidades de usuario incluidas en los tweets cuáles son seguidos por el perfil de estudio. Nuevamente destaca jafurtado, con un 92% de entidades @, que además son seguidos por él.

Los detalles concretos de las entidades se pueden utilizar para realimentar el sistema de forma automática y ampliar la recogida de datos.

4.6. Retweets

En la Tabla VI podemos ver el número de retweets de cada perfil. El mayor porcentaje de retweets son para javiercelaya y futuroslibro.

24symbols	415	66,2%
actualidad_ed	1588	49,6%
anatomiaedicion	804	55,3%
editordigital	115	30,8%
electragir	5	10,6%
futuroslibro	187	71,6%
hubertguillaud	880	43,3%
jacordongarcia	239	29,9%
jafurtado	1155	36,1%
javiercelaya	2360	73,8%
lecturalab	1137	35,5%
libroscm	606	34,6%
librosybitios	1402	61,3%
pubperspectives	1900	59,4%
territorioebook	1399	48,4%

Tabla VI. Nº de retweets y porcentaje de los mismos

	%	@	Autocitas
24symbols	1,4	aitorgrandes angelluengo justohidalgo	Sí
actualidad_ed	4,2	alfaqeque	Sí
anatomiaedicion	2,8	gozque	No
editordigital	1,7	lclibros	No
electragir	0		
futuroslibro	0		
hubertguillaud	2	fbon	Sí
jacordongarcia	2,5	jafurtado	No
jafurtado	0,4	zite	No
javiercelaya	2,5	Joseantvazquez julietalionetti	Sí
lecturalab	2	fgsr	No
libroscm	1,6	libroscm	Sí
librosybitios	2	librosybitios	Sí
pubperspectives	0,3	ebooknewser	Sí
territorioebook	3,1	lecturalab	Sí

Tabla VII. Porcentaje de retweets a seguidos, top 1 de los seguidos y existencia de autocitas

4.6.1. Retweets a seguidos

En la Tabla VII se indica el porcentaje de retweets que se realiza a perfiles seguidos. El número es muy bajo y el que más "cita" es ac-

tualidad_ed. La mayoría se autocita y en algunos casos esa autocita es el top 1.

4.7. Influencia de los retweets

En la Tabla VIII se representa la influencia que tienen los retweets. Valorando la ratio tweet/retweet, los perfiles que tienen mayor influencia son javiercelaya, futuroslibro, 24symbols y librosybitios.

	Tweet/retweet	Total
	0,662	2,448
	0,496	1,532
	0,553	1,568
	0,308	1,550
	0,106	0,128
	0,716	2,149
	0,433	1,164
	0,299	0,845
	0,361	0,830
	0,738	5,441
	0,355	0,918
	0,346	0,925
	0,613	7,371
	0,594	1,374
	0,484	1,464

Tabla VIII. Influencia de los retweets

Cuando además se tiene en cuenta el número total de retweets que tiene un tweet destaca claramente librosybitios y a continuación javiercelaya.

4.8. Nube de etiquetas

En lo que se refiere al análisis de contenido, la nube de etiquetas nos muestra qué palabras son las que se emplean en los tweets destacando tipográficamente la importancia en cuanto al número de apariciones. Una vez dibujada (Figura 7, apéndice), se pueden apreciar las siguientes conclusiones:

- Están más representados los soportes, formatos y plataformas que los contenidos, lo que indica una tendencia clara en las redes sociales. Es decir, se reflejan más los cambios tecnológicos y lo que estos tienen de novedad que los contenidos que alojan. Es un aspecto muy interesante para seguir estudiando.

- García Esparza, S.; O'Mahony, M. P. and Smyth, B. (2012) Mining the Real-Time Web: A Novel Approach to Product Recommendation. *Knowledge-Based Systems*. 29 (2012) 3-11
- Introducción a JSON. <http://www.json.org/json-es.html>
- Java, Akshay, et al. (2009) Why we Twitter: An Analysis of a Microblogging Community. // *Advances in Web Mining and Web Usage Analysis*. 5439 (2009) 118-138. ISSN 0302-9743; 978-3-642-00527-5.
- Lee, K.; Palsetia, D.; Narayanan, R.; Patwary, M. M. A.; Agrawal, A.; Choudhary, A. (2011). Twitter trending topic classification. // 11th IEEE international conference on data mining workshops (ICDMW).
- Narr, S.; De Luca, E. W.; Albayrak, S. (2011) Extracting Semantic Annotations from Twitter. // *Proceedings of the fourth workshop on Exploiting semantic annotations in information retrieval, ESAIR '11*. 15-16.
- The Cocktail Analisis. Informe de resultados Observatorio Redes Sociales. 3ª oleada. Febrero de 2011. <http://www.tcanalysis.com/uploads/2011/02/Observatorio-RedesSociales2011.pdf>
- The Semantic Web: ESWC 2011 Workshops, Revised Selected Papers. , 2012
- Thelwall, Mike; Buckley, Kevanand; Paltoglou, Georgios. (2011). Sentiment in Twitter Events. // *Journal of the American Society for Information Science and Technology*. 62:2 (Feb. 2011) 406-418.
- El uso de Twitter en las editoriales españolas 2011. (2011) Anatomía de la edición. 2011. <http://www.anatomiadelaedicion.com/2011/09/el-uso-de-twitter-en-las-editoriales-espanolas-2011/>
- Yerva, S. R.; Miklós, Z.; Aberer, K. (2012) Quality-Aware Similarity Assessment for Entity Matching in Web Data. // *Information Systems*. 37:4 (2012) 336-351.

Enviado: 2012-04-08. Aceptado: 2012-07-01.
