Generación de un corpus de usuarios basado en divergencias del lenguaje

A.Castellanos[1], J.Cigarran[1,] A.García-Serrano[1]

[1]Universidad Nacional de Educación a Distancia, UNED

{acastellanos,agarcia,juanci}@lsi.uned.es

Abstract. Este artículo presenta el modelado de usuarios sobre la base de los contenidos consultados con anterioridad por éstos. Para generar este modelo se plantea una aproximación basada en divergencias entre términos, en lugar de similitudes. El objetivo es tener un modelado que capture la actividad específica de los usuarios, y no solo aquella más genérica. Este modelado servirá como base para, por ejemplo, el desarrollo de un sistema de recomendación, evitando el problema de sobre-especialización, gracias a la definición más concreta de la actividad de usuario. Se ha creado un corpus de noticias clasificadas y se presenta el corpus de usuarios, de acuerdo a tres perfiles, generado para futuros experimentos en sistemas de recomendación. Se presenta también una comparación que demuestra las ventajas de este modelado frente a otro basado en TF-IDF

Keywords: Modelado de usuarios, divergencia Kulllback-Liebler, Sistemas de Recomendación

1 Introducción

La gran cantidad de contenidos presentes en la web dificulta a los usuarios la identificación de contenidos relevantes. Cada vez con más frecuencia, es necesario el uso de sistemas que le faciliten el descubrimiento de nuevos contenidos. Para intentar solucionar este problema, surgen los sistemas de recomendación. Estos sistemas analizan la actividad previa de los usuarios para detectar sus preferencias y ofrecer contenidos de acuerdo a éstas. Los sistemas de recomendación han sido clasificados tradicionalmente en la literatura en tres categorías [2]:

- Recomendación Basada en Contenidos, que recomienda contenidos similares a aquellos que el usuario ha considerado relevantes, o ha visitado, con anterioridad.
- Recomendación Basada en Filtros Colaborativos, que recomienda contenidos considerados relevantes por usuarios similares al usuario objetivo.
- Recomendación Híbrida que combina el funcionamiento de sistemas que siguen los dos enfoques anteriores.

A pesar de los diferentes enfoques, estos sistemas presentan problemas como la recomendación de contenidos no relevantes o de contenidos poco originales. Los sistemas basados en filtro colaborativo no funcionan correctamente ante la aparición de nuevos usuarios o contenidos de los que no se tiene información (cold-start). Este problema es cubierto por los sistemas basados en contenido que, sin embargo, presentan el problema de sobre-especialización, esto es, recomendación de contenidos muy similares a los que ya han sido considerados por los usuarios, dando lugar a una falta de originalidad.

Se han planteado varias soluciones a la sobre-especialización, algunas de ellas se recogen en [4]. Una de las soluciones propuestas se basa en desarrollar un modelado de los usuarios que recoja mejor sus preferencias. Sobre esta última solución, en [5] se discute sobre diferentes modelados basados en las interacciones previas de los usuarios. A este respecto, en este trabajo se plantea un modelado basado en los contenidos con los que el usuario ha interactuado. Mediante la utilización de divergencia entre términos y más concretamente, la divergencia de Kullback-Liebler, se pretende tener un modelado de usuarios más preciso que mediante técnicas clásicas basadas en similitud y, posteriormente, facilitar la recomendación de contenidos originales.

En los apartados siguientes se incluye una descripción de la colección de noticias, utilizada en los experimentos que se presentan a continuación para el modelado de usuarios, que se han organizado en un corpus con tres perfiles de usuario. Finalmente se incluyen unos comentarios finales y un planteamiento de trabajo futuro.

2 Colección de noticias

Para modelar a los usuarios se ha utilizado como base una colección formada por 5360 noticias en castellano, extraídas de 14 periódicos online, tarea realizada en el marco del proyecto BUSCAMEDIA[1] en 2011. Las noticias estaban clasificadas mediante la anotación de categorías procedente de las secciones de los periódicos. En esta categorización se detectaron una serie de problemas derivados de la falta de normalización. Cada periódico tiene definida su propia taxonomía que, unido a su propio criterio de clasificación propio, dificulta una clasificación global. Por ejemplo, para el mismo tipo de noticias, en algunos periódicos existía una sección llamada Nacional y en otros existía una sección llamada España. Estos problemas hacían que el número de categorías fuese excesivo, aproximadamente unas 400 para las cerca de 5000 noticias.

Por todo ello se decidió desechar esta clasificación y emplear un clasificador que tuviese en cuenta el contenido de la noticia. El clasificador elegido basa su funcionamiento en una aproximación híbrida, descrita en [7], que utiliza la combinación de un clasificador basado en kNN (K-Nearest Neighbor) y un sistema experto basado en reglas. Este clasificador ha sido evaluado frente a los principales clasificadores descritos en la literatura, mejorando los resultados de éstos [7].

[1] http://www.cenitbuscamedia.es/

La clasificación de la colección se ha llevado a cabo en dos niveles, categorías y subcategorías, dando lugar a 19 categorías y 182 subcategorías, empleando los Subject Codes propuestos por la International Press Telecommunications Council (IPTC).

3 Modelado de los Usuarios

Partiendo del cómputo de noticias presentado, el modelado se llevará a cabo a partir del contenido de las noticias consultadas con anterioridad por un usuario. Para ello, se representa al usuario mediante el conjunto de términos más representativo del conjunto de noticias leídas. Se considera un término es relevante si aparece frecuentemente en el conjunto de noticias consultadas y aparece raramente en el resto de la colección.

Para calcular la relevancia de cada término se usa Kullback-Liebler divergence (KLD) [3], que permite ordenar los términos en función de su importancia de acuerdo a la Ecuación 1:

$$KLD_{pD,pC} = pD(t) \cdot \log\left(\frac{pD(t)}{pC(t)}\right) \quad \textbf{(1)}$$

donde $pD(t)$ es la probabilidad de que el término t aparezca en el documento D y $pC(t)$ la probabilidad de que el mismo término t aparezca en el resto de la colección.

Mediante la utilización de KLD, se espera descubrir aquellos términos que sirvan para diferenciar a un usuario del resto, generando así un modelo que facilite la recomendación de contenidos originales. Se ha optado por el uso de KLD frente a otras técnicas clásicas como TF-IDF ya que se ha probado de manera experimental [1] que, para categorización de textos, el uso de técnicas probabilísticas como KLD ofrece mejores resultados que técnicas heurísticas como TF-IDF. Esto es debido a qué, a la hora de otorgar pesos a los términos, KLD tiene en cuenta la información de la clase en la que se encuentran, mientras que TF-IDF no. Estos resultados son extrapolables a este trabajo, considerando a los modelos de usuario como clases en las que clasificar los contenidos susceptibles de ser recomendados.

Siguiendo este enfoque, en [6], se presenta un modelado de usuarios similar al propuesto en este trabajo. Dicho modelado se evalúa frente a otras técnicas de modelado tradicionales (como TF-IDF), mejorando considerablemente los resultados. En el trabajo citado anteriormente, se presenta una propuesta para modelar las preferencias específicas de un usuario frente a las del resto de los usuarios. En contraste, nuestra aproximación utiliza únicamente el contenido consultado previamente por un usuario, sin tener en cuenta al resto de usuarios. Con ello se evita la influencia de la actividad de otros usuarios en el modelado.

De cara a cubrir diferentes comportamientos de usuarios a la hora de consultar noticias, se ha decidido crear 3 perfiles: El primero de ellos se corresponde con un usuario genérico que no esté interesado en ninguna temática concreta y consulte noticias de todas las categorías. Para simular a un usuario que esté interesado en una temática concreta se ha desarrollado otro perfil de usuario. Este perfil consulta sólo noticias de una única categoría. Por último se ha simulado un usuario interesado en una temática aún más reducida que en el perfil anterior. Éste último consulta solo noticias de una

única subcategoría. Para cada perfil se han generado un total de 1000 usuarios sobre los que se ha aplicado el modelado propuesto en el punto anterior.

En las tablas 1,2 y 3 se muestra un ejemplo de los 10 términos de más peso para los modelos asociados a cada uno de los perfiles de usuario mediante la aplicación de KLD. En las Tablas 4,5 y 6 se muestran los términos de más peso para los mismos perfiles mediante la aplicación de TF-IDF. Comparando los resultados obtenidos mediante KLD y TF-IDF se puede observar que, a medida que el modelo es más concreto (i.e. existe menor número de clases) los resultados de TF-IDF y de KLD tienden a equipararse. Mientras que en modelado del usuario genérico únicamente comparte un término (cardenal), en el modelado del usuario centrado en una subcategoría se comparten 8 de los 10 términos. Estos resultados confirman los presentados en [1], confirmando que el hecho de que KLD recoja la información de clase hace que su uso sea más interesante como técnica de modelado que TF-IDF. En los resultados obtenidos también se puede observar que con TF-IDF aparecen términos genéricos en el modelo, los cuales aportan poca información (e.g. noticia, 2010 o modera). En los resultados obtenidos con KLD, por el contrario, los términos identificados son más significativos de manera general.

Teniendo solo en cuenta KLD y comparando los modelos generados para cada perfil, se puede observar que los términos del usuario genérico guardan menos relación entre sí que los del usuario interesado en secciones relacionadas con las categorías. Esto es debido a que la temática, y por extensión el vocabulario, es más concreta en el caso del segundo perfil. Este fenómeno se confirma en el perfil del usuario interesado en secciones relacionadas con las subcategorías (i.e. un usuario cuyos intereses son mucho más específicos). Teniendo en cuenta los pesos asociados a los términos, se puede observar que, a medida que el perfil del usuario se enfoca en una temática más concreta, los pesos son más altos y por tanto los términos son más representativos. Estos resultados confirman que, cuanto más centradas estén las preferencias de un usuario, más representativa será la representación que se tenga de este.

Término	Peso
descenso	**0.009155**
constitucional	**0.006711**
climatico	**0.006328**
alejamiento	**0.006155**
estepa	**0.005740**
benedicto	**0.005629**
debil	**0.005476**
cardenal	**0.005091**
baix	**0.004840**
cambio	**0.004569**

Tabla 1. Términos KLD Usuario genérico

Término	Peso
euros	**0.016727**
alquiler	**0.008365**
datos	**0.007353**
china	**0.007292**
electrico	**0.006877**
agosto	**0.006611**
credito	**0.006521**
climatico	**0.006352**
aol	**0.006219**
banco	**0.005699**

Tabla 2. Términos KLD Usuario categoría

Término	Peso
garofalo	**0.024684**
mafia	**0.024675**
narcotrafico	**0.020647**
acido	**0.019549**
cosco	**0.014470**
cocaina	**0.012388**
ndrangheta	**0.011065**
dinero	**0.010325**
empresarios	**0.010292**
disolvio	**0.010214**

Tabla 3. Términos KLD Usuario subcategoría

Término	Peso
nuboso	**0.018208**
universo	**0.016213**
noticia	**0.015676**
viento	**0.015638**
galaxia	**0.015572**
cardenal	**0.013471**
modera	**0.013196**
premio	**0.012887**
2010	**0.012503**
intervalo	**0.012304**

Tabla 4. Términos TFIDF Usuario genérico

Término	Peso
millon	**0.035908**
euros	**0.0266840**
precio	**0.0227103**
mineria	**0.0206410**
vivienda	**0.0200239**
banco	**0.0168816**
electrico	**0.0168733**
noticia	**0.01627789**
mercado	**0.0155405**
bce	**0.0146306**

Tabla 5. Términos TFIDF Usuario categoría

Término	Peso
garofalo	**0.070150**
mafia	**0.058134**
acido	**0.050663**
narcotrafico	**0.047870**
cosco	**0.041123**
cocaina	**0.032854**
ndrangheta	**0.031447**
disolvio	**0.029028**
blanqueo	**0.027392**
bolivia	**0.026580**

Tabla 6. Términos TFIDF Usuario subcategoría

4 Comentarios finales

Se ha desarrollado un modelado de la actividad de los usuarios basado en el contenido previamente consultado por éstos. Para ello se ha utilizado una aproximación basada en divergencia de términos. Mediante esta aproximación es posible identificar aquellos términos que son más significativos, permitiendo describir de manera efectiva la actividad previa del usuario. Se han desarrollado distintos perfiles reflejando diferentes comportamientos de los usuarios. Estos perfiles permiten comprobar la adaptabilidad del modelado propuesto a diferentes situaciones de funcionamiento. En conjunto con este modelado, se ha generado una colección de noticias que, posteriormente, se ha clasificado en dos niveles.

Basándose en el método de evaluación presentado en [6], se plantea un trabajo futuro de evaluación del modelado, utilizando para ello la información de clase que se tiene de cada una de las noticias. Gracias a estar información, se podrá comprobar si los términos del modelo de un usuario se corresponden con la categoría de las noticias que este ha consultado.

El modelado de usuarios y la colección de prueba (corpus de noticias y de usuarios de tres perfiles) presentado en este trabajo pretenden ser utilizados como base para un futuro sistema de recomendación de contenidos que facilite la identificación de, no solo las preferencias generales del usuario, sino también de aquellas más específicas. De esta manera se espera poder resolver el problema de sobre-especialización.

Agradecimientos. Este trabajo ha sido parcialmente financiado por el Gobierno Regional de la Comunidad de Madrid mediante la Red de Investigación MA2VIRMR (S2009/TIC-1542) y por el Gobierno Español mediante el proyecto BUSCAMEDIA (CEN-20091026).

Referencias

1. Bigi, B.: Using Kullback-Leibler Distance for Text Categorization, In *Proceedings of the ECIR-2003, volume 2633 of Lecture Notes in Computer Science*, pages 305-319, 2003.
2. Candillier, L. Jack, K. Fessant, F. Meyer, F.: State-of-the-art recommender systems. *Collaborative and Social Information Retrieval and Access-Techniques for Improved User Modeling*, pages 1–22, 2009.
3. Kullback, S. Leibler. R.A.: On information and sufficiency. *Annals of Mathematical Statistics*, *22(1)*, pages 79-86, 1951.
4. Lops, P. Gemmis, M. Semeraro, G.: Content-based Recommender Systems: State of the Art an Trends. In *Recommender Systems Handbook*, F. Ricci, L. Rokach, B. Shapira, P.B.Kantor, Eds. Springer US, Boston, MA, pages 73–106. 2011.
5. Pazzani, M. Billsus, D.: Content-Based Recommendation Systems. *The Adaptive Web: Methods and Strategies of Web Personalization. Vol. 4321 of LNCS*, pages 325-341. 2007.
6. Shmueli-Scheuer, M. Roitman, H. Carmel, D. Mass, Y. Konopnicki, D.: Extracting User Profiles from Large Scale Data. In *Proceedings of the 2010 Workshop on Massive Data Analytics on the Cloud*, 2010.
7. Villena-Román, J. Collada-Pérez, S. Lana-Serrano, S. González-Cristóbal, J. Hybrid Approach Combining Machine Learning and a Rule-Based Expert System for Text Categorization, In *Proceedings of the Twenty-Fourth International Florida Artificial Intelligence Research Society Conference*, 2011.