Sistemas de Recomendación basados en Lenguaje Natural: opiniones vs. valoraciones

John A. Roberto[1], Ma. Antònia Martí[1], Paolo Rosso[2]
[1] Dpto. de Lingüística, Universidad de Barcelona,
Gran Via de les Corts Catalanes, 585. 08007 Barcelona, España
[2] NLE Lab. - ELiRF, DSIC, Universidad Politécnica de Valencia
{roberto.john, amarti}@ub.edu, prosso@dsic.upv.es

Resumen. La construcción de los perfiles de usuario es una de las fases más críticas en el desarrollo de los Sistemas de Recomendación. Actualmente, la mayoría de estos sistemas construyen los perfiles de usuario siguiendo modelos explícitos o implícitos de adquisición de datos. En oposición a estos modelos clásicos, nuestra investigación se centrará en la construcción de perfiles a partir de las opiniones de los usuarios expresadas en lenguaje natural.

Palabras Clave: Sistemas de Recomendación, perfiles de usuario, PLN, opiniones de usuario.

Descripción del trabajo

Los Sistemas de Recomendación (SR) son un tipo particular de aplicaciones y de técnicas especialmente desarrolladas para filtrar la información. Los SR trabajan efectuando predicciones sobre un ítem o conjunto de ítems que podrían ser de interés para un usuario particular. Estos sistemas están enfocados a la eliminación de ítems irrelevantes del flujo de datos. Los SR basan su funcionamiento en el conocimiento que tienen sobre las preferencias de los usuarios y que es almacenado en los perfiles de usuario. La construcción de los perfiles de usuario es una de las fases más críticas del desarrollo de los SR y ahí centraremos nuestra investigación.

En la actualidad, la mayoría de los SR construyen los perfiles de usuario según dos modelos [1] [6] [5]: preguntando directamente a las personas cuales son sus preferencias (modelo explícito) o infiriéndolo de sus acciones (modelo implícito). La ponderación de un ítem en base a una escala ordinal o cualitativa es típica del primer modelo mientras que la selección de un ítem o el tiempo invertido es su visualización, es característica del segundo.

En general, la lógica que gobierna ambos modelos es muy elemental y por lo tanto impide adquirir información compleja. Los modelos explícitos exigen al usuario un esfuerzo cognitivo importante [3] pues, dependiendo el dominio, las personas pueden no estar cualificadas para valorar un producto. Los modelos implícitos, por su parte, tienen limitaciones a la hora de interpretar algunas conductas ambiguas [5]; esto hace

que no sean bien recibidos por los usuarios quienes muchas veces no comprenden el motivo de la recomendación.

En oposición a los modelos clásicos, nuestra investigación se centrará en la construcción de perfiles a partir de las opiniones en lenguaje natural. El empleo del lenguaje natural para la creación de perfiles permite adquirir información compleja [11] [12] [9], que va más allá de la simple ponderación o la selección de un ítem. Adicionalmente, esta técnica mejora la forma en que el usuario se comunica con el sistema puesto que las opiniones, a diferencia de las puntuaciones, no exigen un conocimiento en profundidad de los ítems. El resultado es, pues, una reducción significativa del esfuerzo por parte del usuario y el potencial incremento en la calidad de los datos.

En tareas de recomendación la información lingüística se ha empleado básicamente para la representación del contenido textual de los ítems pero solo en contados casos se ha hecho servir a modo de estrategia de retroalimentación. Los MLD[1] [7] [10], por ejemplo, permiten que los usuarios expresen las valoraciones con palabras ("bonito", "malo", etc.) en lugar de valores numéricos y los SR Conversacionales (SRC) [4] formulan preguntas que los usuarios deben responder ordenadamente para ir refinando la recomendación de forma gradual. Una limitación importante de estos sistemas es que no trabajan con texto libre y para que funcionen tienen que imponer algún tipo de restricción sobre el lenguaje que utiliza el usuario. Nuestra investigación pretende arrojar luz en este sentido, ofreciendo alternativas al tratamiento de texto libre para crear y mantener los perfiles de usuario.

Para crear los perfiles de usuario a partir de opiniones en lenguaje natural utilizaremos estereotipos[2] basados en la forma en que diferentes grupos de usuarios tienen de opinar. Un estereotipo, según nuestro modelo, se compone de un conjunto de características lingüísticas que evidencian la forma de opinar de un grupo y las preferencias asociadas al grupo. De esta manera el proceso de recomendación consiste en hacer que el usuario activo[3] herede las preferencias de los usuarios que se expresen de modo similar.

Proponemos dos métodos alternativos para la obtención de los rasgos lingüísticos que definirán cada estereotipo. El primero radica en el análisis lingüístico de un conjunto significativo de críticas de productos agrupados según las puntuaciones asignadas por los usuarios para dichos productos. La fuente de datos que emplearemos será HOpinion (http://clic.ub.edu), un corpus en castellano de críticas de hoteles recuperadas de TripAdvisor[4]. El corpus tiene un total de 4750 textos (cerca de un millón de palabras) de críticas de hoteles, con una medida promedio de 135

[1] SR basados en Modelos Lingüísticos Difusos.

[2] Los estereotipos son mecanismos empleados para efectuar descripciones parciales de situaciones que se suceden frecuentemente. En los SR los estereotipos se emplean para asignar un usuario a un grupo de usuarios similares del que hereda sus preferencias.

[3] Usuario que es objeto del proceso de recomendación actual.

[4] TripAdvisor es la mayor comunidad de viajeros *on line* del mundo y cuentan con más de 20 millones de opiniones sobre establecimientos y destinos basadas en la experiencia personal de sus usuarios.

palabras por crítica. La valoración de las críticas va en una escala del 10 al 50 en la siguiente relación: 10 / pésimo, 20 / malo, 30 / normal, 40 / muy bueno y 50 / excelente. HOpinion ha sido anotado con información morfosintáctica y corregido manualmente. Además, se le ha aplicado un *chunking* para identificar los constituyentes básicos.

El análisis lingüístico de estas críticas busca caracterizar los usuarios por su forma de utilizar el lenguaje, de manera que se puedan agrupar en una tipología de registros: culto (CU), coloquial (CO) y neutro (NE). Asumimos, por tanto, una relación directa entre la forma de expresarse (registro) y el perfil del usuario. En el análisis lingüístico se aplicarán técnicas de PLN tales como el análisis morfosintáctico y el *chunking* empleando las herramientas disponibles. Adicionalmente y valiéndonos de la información obtenida en los procedimientos anteriores, se prevé un estudio más superficial orientado a la búsqueda de segmentos recurrentes (n-gramas). Los datos resultantes (léxico, *chunks*, expresiones, etc.) serán clasificados manualmente para asignarles el registro.

El segundo método para obtener los patrones consiste en anotar las criticas de hoteles del mismo corpus con nuestra tipología de registros (CU, CO, NE) y, de la misma manera que se hizo en el procedimiento anterior, analizar lingüísticamente los textos para obtener los patrones asociados a cada registro. Un punto que puede parecer delicado de esta aproximación es, justamente, la razón que nos lleva a asociar cada texto de opinión a un registro. No obstante, creemos que se trata de una inferencia equivalente a la que se suele asumir en la literatura sobre el análisis de sentimiento cuando se hacer corresponder la valoración vertida por el usuario en la crítica con la valoración numérica global [8]. La coincidencia o no de los patrones obtenidos en los dos procedimientos nos servirá para constatar el grado de fiabilidad que podemos atribuir a dichos patrones para discriminar los textos de opinión por registro.

Como resultado de aplicar esta metodología se dispondrá de un recurso compuesto por un léxico de palabras, frases y expresiones que nos servirá para predecir el registro al que pertenece un determinado texto de opinión. Para evaluar su capacidad de predicción en otros dominios, clasificaremos de forma automática un conjunto de críticas de películas del corpus MuchoCine[5] [2] y compararemos los resultados con las etiquetas asignadas de forma manual por un grupo de anotadores. Las conclusiones obtenidas con esta investigación nos servirán, entre otras cosas, para determinar si es viable ampliar la tipología de registros propuesta a otros de diferente matiz: directo / indirecto, experto / inexperto, etc., lo que supondría contar con un número mayor de perfiles de usuarios.

[5] El corpus de MuchoCine (http://www.lsi.us.es/~fermin/corpusCine.zip) tiene un total de 3.878 críticas y aproximadamente 2 millones de palabras, con una media de 546 palabras por crítica. Cada crítica ha sido procesada con FreeLing para obtener información léxica, morfosintáctica y semántica codificada en diferentes ficheros.

Agradecimientos. Los autores agradecen a los proyectos de investigación: MICINN TEXT-ENTERPRISE 2.0 TIN2009-13391-C04-03 y TEXT-KNOWLEDGE 2.0. TIN2009-13391-C04-04 (Plan I+D+i).

Referencias

[1] Brusilovsky, P. y Maybury, M. From Adaptive Hypermedia to the Adaptive Web. *Communications of the ACM*. 45(5): 31-33, 2002.

[2] Cruz Mate, Fermín, et al. Clasificación de documentos basada en la opinión: experimentos con un corpus de críticas de cine en español. *Procesamiento del lenguaje natural*. N. 41 (sept. 2008). ISSN 1135-5948, pp. 73-80

[3] De Gemmis, M., Iaquinta, L., Lops, P., Musto, C., Narducci, F. y Semeraro, G.. Preference learning in recommender systems. *Preference Learning (PL-09) ECML/PKDD-09 Workshop*, 2009.

[4] Derek, Bridge. *Towards conversational recommender systems: A dialogue grammar approach*. Proceedings of the Workshop in Mixed-Initiative Case-Based Reasoning, 2002.

[5] Kelly, D. Implicit feedback: Using behavior to infer relevance. A. Spink and C. Cole (Eds.) *New Directions in Cognitive Information Retrieval*. Springer Publishing: Netherlands. 2005: 169-186.

[6] Montaner, M., López, B. y J. L. D. L. Rosa. A taxonomy of recommender agents on the internet. *Artificial Intelligence Review*, 19(4):285-330, 2003.

[7] Morales-del-Castillo, J.M., Herrera-Viedma, E. Peis. *Modelo semántico-difuso de un sistema de recomendaciones de información para bibliotecas digitales universitarias*. II Simposio sobre Lógica Fuzzy y Soft Computing (LFSC 2007). pages. 73–80, 2007.

[8] Moreno Ortiz, A., Pineda Castillo, F. y Hidalgo García, R. *Análisis de Valoraciones de Usuario de Hoteles con Sentitext: un sistema de análisis de sentimiento independiente del dominio* . Procesamiento del Lenguaje Natural, Revista no45, septiembre 2010, pp 31-39.

[9] Reitter, D., Covaci, S., Oltean, F., Bacanu, C. y Serbanuta, T. Hybrid natural language processing in a customer-care environment. *Proc. of the 11th TaCoS*, 2001.

[10] Sáchez, Pedro José. *Modelos para la combinación de preferencias en toma de decisiones: herramientas y aplicaciones*. PhD thesis, Departamento de Ciencias de la Computación e Inteligencia Artificial. Universidad de Granada. 2007.

[11] Schmitt S. y Bergmann R. A Formal Approach to Dialogs with Online Customers. *The 14th Bled Electronic Commerce Conference*. Bled, Slovenia, 2001:309-28.

[12] Zukerman, I. y Litman, D. Natural Language Processing and User Modeling: Synergies and Limitations. *User Modeling and User-Adapted Interaction*. Vol. 11: 129-158, 2001