Modelado y propagación de valores de sentimiento en relaciones de usuario

Ramón Rivera Camacho, Ricardo Barrón Fernández, Adolfo Guzmán Arenas

Instituto Politécnico Nacional, Centro de Investigación en Computación, México

Resumen. Se propone un método de modelado de relaciones personales basado en dos aspectos principales, la relevancia y componente emocional subjetiva del usuario y la categorización objetiva de las relaciones. Se presenta dicho modelo como un grafo conexo con la suma de todos los componentes como nodo principal rodeado de relaciones paradigma y relaciones descubiertas conectadas a través de niveles de relevancia. En segunda instancia se describe la manera de propagar el componente sentimental a través del modelo construido por medio del método de retropropagación (*back-propagation*). Finalmente se presenta las propiedades de dicho modelo, resultados obtenidos y posibles usos del mismo.

Palabras clave: Análisis de sentimientos, similitud semántica, aprendizaje automático.

Modelling and Propagation of Sentiment Values in Relations between Users

Abstract. A method of modeling personal relationships is proposed grounded on two main aspects, the subjective relevance and emotional component of the user, and the objective categorization of their relationships. This model is built as a connected graph with its principal node depicting the complete sum of all of the emotional components surrounded by paradigm and discovered relationships joined to this according to their level of relevance. As a compliment, a way to spread the sentimental component through the model by back-propagation method is described. Finally, some properties of the model are shown, as well as obtained results and possible usages of this.

Keywords: Sentiment analysis, semantic similarity, machine learning.

1. Introducción

A medida que la cantidad de información textual incrementa dentro de la red, el sector dedicado a la minería de opiniones, también conocido como análisis de sentimiento, obtiene mayor importancia, ya que en este recaen las tareas de obtención de características y tópicos, así como la clasificación de las preferencias de usuario para éstos, a menudo obtenidos desde redes sociales, blogs de opiniones, entre otros; sin

embargo, poco desarrollo se ha realizado para analizar sentimientos que afectan al usuario en sí, así como el efecto en sus relaciones.

El análisis de sentimientos principalmente está enfocado a reportar la actitud de un sujeto con respecto a un objeto determinado, esta tarea puede realizarse, en general, de dos maneras, supervisada y no supervisada, siendo la primera la más precisa debido a que los clasificadores son entrenados con un conjunto de datos representativos, el corpus. Pueden destacarse dentro de este ramo las técnicas de aprendizaje automático (*Machine learning*), como el probabilístico Naïve Bayes que es un método relativamente sencillo pero que se caracteriza por realizar una buena clasificación de sentimientos, sobre todo cuando las características son altamente dependientes [2-3], y el basado en hiperplanos de categorización, Máquinas de Soporte Vectorial (*Support Vector Machines*, *SVM* por sus siglas del inglés), siendo este último el que mejor desempeño reporta [6].

En contraparte, los métodos no supervisados exhiben un menor desempeño pero minimizan el trabajo de clasificación previa y de dependencia de dominio, algunos de ellos trabajando por medio de palabras semilla (*Seed words*) y calculando la orientación semántica de las frases [6-8].

Conforme al aumento en la cantidad de datos es necesario que el análisis tenga una reducción en la dimensión de búsqueda, es decir, la extracción de tópicos. Mei, et al., asume la existencia de una mezcla de ellos en el documento y que por lo tanto pueden ser reducidos utilizando una distribución de probabilidad [5], mientras que Blei, et al. [4], se basa en el principio de intercambiabilidad de De Finetti para construir el modelo probabilístico. En conclusión, ambos basan su funcionamiento en la construcción de modelos probabilísticos generativos cuya tarea es encontrar los tópicos latentes ocultos dentro del documento.

Siguiendo la misma línea, además de utilizar las palabras semilla, pero enfocado a un escenario más general Lin et al. [1] proponen un conjunto de palabras con contenido sentimental conocido, la palabras paradigma, para clasificación de sentimientos no supervisada a nivel de documento, la cual realiza la extracción de tópicos por medio de clasificadores bayesianos, sin embargo, al igual que Blei [4] y Mei [5], el modelo representa los resultados como bolsas de palabras (*bag-of-words*), que no presentan relación entre ellos.

Sin embargo, la falta de relación entre resultados presenta inconsistencia para la interpretación de éstos, necesitando otro nivel de clasificación extra si el problema requiere de éste tipo de información. Blei, et al. [9], retoma la idea del proceso estocástico de restaurante chino (*Chinese restaurant process*), para inferir la jerarquía de los datos a analizar. Por otro lado, para lidiar con este problema, Kim, et al. [10], basándose en la idea que un documento presenta aspectos que pueden ser organizados de manera natural, construyen un árbol de aspectos-sentimientos que describen dicha relación.

En el presente artículo se considera la etapa final del problema de describir el cambio de humor del usuario por medio de valores de sentimiento dentro del conjunto de relaciones de usuario.

Se retoman las ideas previamente tratadas para realizar el modelado del estado de humor del usuario, asumiendo por definición las relaciones de parentesco del mismo, las relaciones paradigma, y considerando la afectación sentimental por medio de la propagación de componentes sentimentales. La descripción se centra en la construcción

de un modelo jerárquico por medio de distancias semánticas y la probabilidad de pertenecer a cierto conjunto superior, además, se propone un método de propagación del componente sentimental hasta alcanzar el nodo raíz por medio de técnicas bayesianas.

2. Análisis

Se aborda, como primer acercamiento, el modelado del cambio emocional debido a factores externos, con la construcción de un grafo que satisfaga la siguiente idea sencilla: "La descripción sentimental del usuario parte de la afectación de tópicos externos y mientras éste se refiera más a un tópico en específico, más relevancia sentimental tendrá y más cambios realizará en el usuario".

Para la implementación de esta idea se construirá de manera no supervisada un modelo jerárquico que contenga las relaciones de parentesco del usuario, a partir de un conjunto de relaciones base que se complementarán con base en la similitud de nuevos términos, posteriormente, el valor de sentimiento que contenga se propagará hasta llegar al nodo principal tomando en cuenta la relevancia del nodo padre, obtenida por medio de la cantidad de veces que el tema haya sido tratado previamente. Este proceso es dividido en 2 etapas, la complementación y construcción del grafo que describe las relaciones y la propagación de las componentes de sentimiento.

Para crear el modelo de relaciones de usuario (URM), se utiliza un grafo no dirigido que contiene como nodos al conjunto de valores de relaciones conocidos, las relaciones paradigma (*PRs*), dadas por definición y que además presentan la propiedad es-parte-de (*part-of*) o pertenencia (∈) con respecto al conjunto que define su nodo padre [11], por ejemplo:

Padre(s) es-parte-de Familia(s) → Padre(s) ∈ Familia(s)
Novio(s) es-parte-de Conocido(s) → Novio(s) ∈ Conocidos(s)

Donde, los nodos Familia y Conocidos mantiene la misma relación con un padre y éstos a su vez repiten el proceso con nodos de orden superior.

En primera instancia se debe partir de un nodo principal que define el total de las contribuciones de componentes sentimentales y el punto de partida para realizar la búsqueda dentro de las relaciones paradigma, el nodo "TODO" ("*ALL*").

Por otro lado, aquellos términos que presenten valores de sentimiento y relevancia significantes, pero que no estén presentes dentro del URM serán clasificados tomando en cuenta distancias semánticas que evalúan el parentesco de dos conceptos evaluando los enlaces semánticos es-un (*is-a*) y parte-de (*part-of*) utilizando parámetros como la distancia más corta entre conceptos o la distancia a un concepto superior común (*subsumer*), en lo posterior tratado como mscs (Del inglés, *the most specific common subsumer*).

Se propone utilizar la medida de similitud semántica de Wu-Palmer [12], descrita en (1).

$$sim_{wp}(c_i, c_l) = \frac{2 * \min_{p \in pths(mscs(c_i,c_j),rt)} len_e(p)}{\min_{p \in pths(c_i,c_j)} len_e(p) + 2 * \min_{p \in pths(mscs(c_i,c_j),rt)} len_e(p)} \quad (1)$$

donde:

rt: Raíz ontológica
pths(x, y): Conjunto de caminos entre los conceptos x e y.
$len_e(x)$:Cantidad de aristas hacia el camino x.

A diferencia de otras medidas de similitud semántica, Wu-Palmer toma en cuenta la distancia al mscs y no presenta necesidad de un corpus de entrenamiento, como otras presentadas en la literatura [12-15]. Realizando una búsqueda por anchura (BFS) y comparando esta similitud con los de un umbral definido por el usuario, para decidir la pertenencia a uno o varios nodos padre, generando las relaciones orden N (Nth-order discovered relationships, NDORs). Como se muestra en la figura 1:

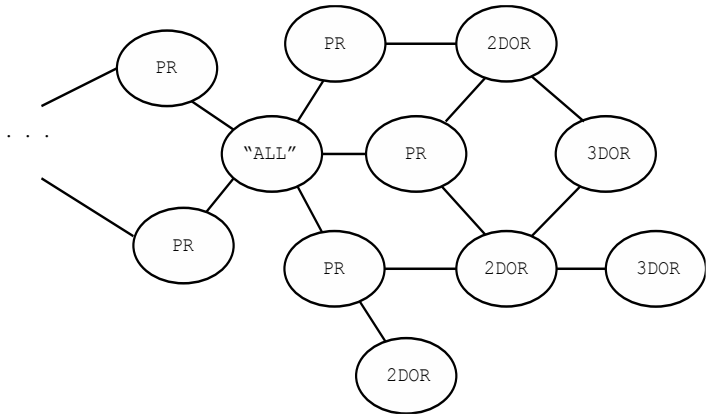

Fig. 1. Posible distribución del URM, se muestra el nodo principal "ALL", rodeado de relaciones paradigma (PR) y relaciones descubiertas de orden N (NDOR).

Para realizar el cálculo de la componente emocional adjunta a la relación, utilizamos el método de propagación hacia atrás (*backpropagation*) del valor de sentimiento V_s considerando la relación jerárquica y semántica del portador con los nodos relacionados de orden superior, como se describe:

1. Sea un C_j la probabilidad de propagar un valor de sentimiento N_{act} a través de un conjunto finito de relaciones jerárquicas $N_r = \{N_{r1}, N_{r2}, \ldots, N_{rn}\}$ de orden superior alcanzables, dada por:

$$P(C_j) = P(N_{act}|N_r) \tag{2}$$

$$P(N_p|N_{rel}) = \frac{P(N_{rel}|N_p)P(N_p)}{P(N_{rel})} \tag{3}$$

$$P(N_{rel}) = P(N_{os}|N_r) = \frac{P(N_r|N_{os})P(N_{os})}{N_r} \tag{4}$$

donde:

N_{os} = Nodos de orden superior, compuestos por los nodos directos, N_d, y los nodos alcanzables N_r.

N_{act} = Nodos actuales, compuestos por el nodo de propagación N_p y los nodos relacionados, N_{rel}.

2. Sea C_s la razón contribución semántica del portador del componente sentimental, N_{act}, con respecto al conjunto finito de relaciones directas de orden superior Np = {N_{p1}, N_{p2}, …N_{pn}}, por lo tanto:

$$C_s = C_{s1} + C_{s2} + \cdots + C_{sn-1} \tag{5}$$

con:

$$C_{si} = \frac{N_{act}|N_{pi}}{N_{act}|N_{p1} + N_{act}|N_{p2} + \cdots + N_{act}|N_{pn}} \tag{6}$$

donde:

N_{act} = Nodo portador de la componente sentimental.
N_{p1}, N_{p2},…,N_{pn} = Nodos relacionados directamente con N_{act}.
$N_{act|Np}$ = Razón de valor semántico entre N_{act} y N_p.

Finalmente, la propagación se dará sucesivamente por medio de la función F_p, previamente descrita, dada por:

$$F_p \rightarrow V_S * C_s * C_j \tag{7}$$

Por lo tanto, el proceso se resume como un método de "complementación-propagación" de N número de objetos sentimiento-relación; asumiendo un conjunto de relaciones candidato (PRs), con una relevancia suficiente y conteniendo tanto el valor de componente sentimental a propagar como la relación adjunta, se propone el siguiente algoritmo de construcción del URM.

```
URM           (Modelo de relaciones de usuario)
PRs           (Conjunto de candidatos a agregarse al MRU)
Sim-umbral    (Umbral de similitud)
robjs         (Objetos de relación)
parents       (Nodos padre)

sub add_to_URM(robjs)
parents = []

for each robj in robjs
    if robj in URM
        URM[robj]->count = URM[robj]->count + 1
        backpropagation (URM[robj]);
    else
        if robj in PRs
            PRs[robj]->count = PRs[robj]->count + 1

            if PRs[robj] >= sim-umbral;
                parents = calculate_distances()
                attach_parents(PRs[robj], parents)
                backpropagation(PRs[robj])
```

```
            end if
        else
             add(robj, PRs)
             PRs[robj]->count = 1
        end if
     end if
end for
```

donde las rutinas *attach_parents(PRs[obj], parents)* y *backpropagation()*, simplemente realizan el enlace con los nodos de orden superior y la propagación del componente de sentimiento como fue descrito respectivamente y *add(robj, PRs)* solamente agrega el objeto de entrada a las relaciones candidato.

3. Resultados

La presente sección se enfocará a analizar los casos que se presentan en la propagación del componente sentimental, cabe mencionar que no está dentro del alcance la obtención de los objetos de sentimiento, los cuales están definidos como las tuplas (n, p, v), que representan (nombre, POS, valor).

Para tener control sobre la disipación de la componente en cada escenario, se propone un URM compuesto solamente por relaciones paradigma sin relevancia previa.

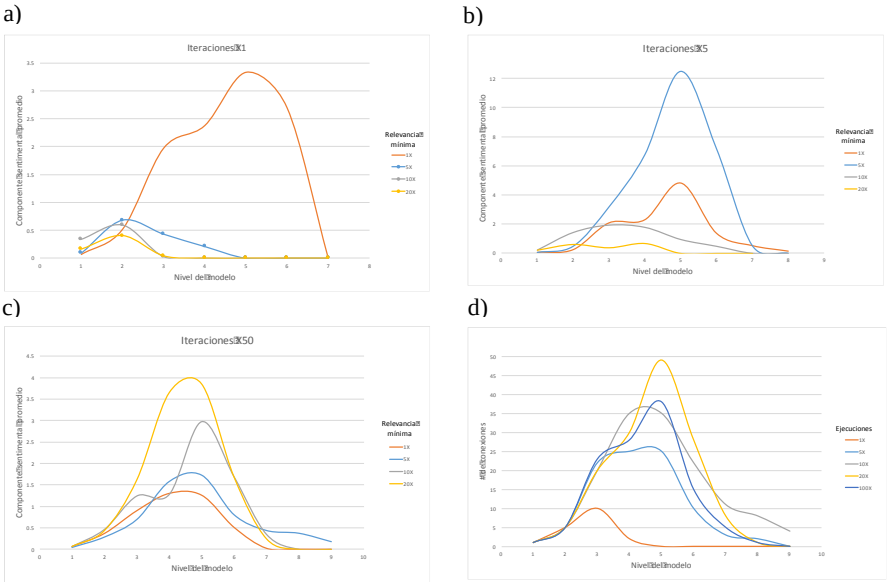

Fig. 2. Componente sentimental promedio a través de los niveles del URM para (a) i={1iX }y r={1rX, 5rX, 10rX, 20rX},(b) i={5iX}y r={1rX, 5rX, 10rX, 20rX}, (c) i={50iX}y r={1rX, 5rX, 10rX, 20rX}. (d) Número de relaciones por nivel para una relevancia fija r = {5rX} e i= {1iX, 5iX, 10iX, 20iX, 100iX}.

Además se provee de un pool de objetos de sentimiento de prueba vagamente clasificados con pertenencia a las principales relaciones paradigma adjuntas al nodo principal, éstos objetos son los que se espera generen una componente sentimental con su correspondiente propagación dentro del modelo, aunados a ellos se suplen un conjunto de objetos aleatorios con componente sentimental pero que no son compatibles con el URM.

Se estipulan valores base para cantidad de iteraciones (iX) y valores de relevancia (rX) y se desarrollan pruebas con valores múltiplo de éstos para los siguientes escenarios, con las especificaciones antes descritas:

1. Tuplas (n, p, v), donde: (n ∈URM, v∈[-3 , 3]), ejecuciones de prueba i={1iX, 5iX, 50iX}, graficados para relevancia r={1rX, 5rX, 10rX, 20rX}
2. Tuplas (N, p, v), donde: N = {n, $\bar{n}$} |{n ∈ URM, $\bar{n}$ ∉URM, $|n| = |\bar{n}|$}, v∈[-2 , 2]), ejecuciones de prueba i={3iX, 5iX, 10iX} con relevancia fija de r = 10rX.

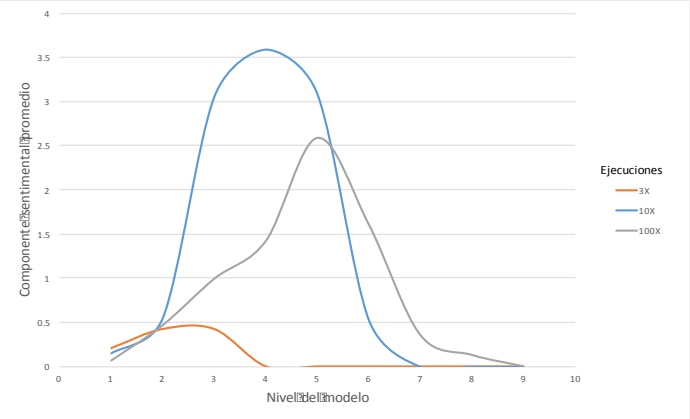

Fig. 3. Componente sentimental promedio para N = {n, $\bar{n}$} | {n ∈ URM, $\bar{n}$ ∉URM, $|n| = |\bar{n}|$} con i= {3iX, 5iX, 10iX} y r = 10rX.

De los resultados se señalan los siguientes puntos:

1. La distribución de los valores de sentimiento a través de los niveles del modelo es similar independientemente de la cantidad de ejecuciones.
2. La cantidad de niveles aumenta con respecto se ejecute el modelo, de igual manera, a medida que se aumenten las ejecuciones la relación de candidatos debe tender a cero.
3. En niveles centrales del grafo se encontrarán el mayor número de conexiones y, por ende, también se presentará mayor relevancia en estos niveles
4. El parámetro de relevancia de un objeto de sentimiento impacta directamente en la construcción del modelo de relaciones, necesitando mayor cantidad de ejecuciones.

En los resultados presentados las relaciones son seleccionadas de manera aleatoria, por lo cual el modelo debería complementarse en anchura, sin embargo, en escenarios de usuario en los cuales se tratan con más cantidad menos tópicos, los niveles no

crecerán de manera tan uniforme, por lo tanto se podría ver con más claridad una tendencia en cuanto a tópicos con mayor relevancia subjetiva para el usuario.

Además, ya que se tiene un grafo conexo cuyo nodo principal es la representación del total de la suma de relaciones de usuario, dada una relación aleatoria es posible obtener la ruta hacia este nodo principal respecto a la relevancia de las relaciones, es decir, obtener el camino más probable de afectación sentimental de la relación.

4. Conclusiones y trabajo futuro

En este artículo, se define un primer enfoque al problema de modelar el cambio de humor debido a factores externos de un usuario. Se propone la construcción de un modelo jerárquico cuyos nodos exhiban la propiedad parte-de (part-of) con el fin de mantener una conexión entre las relaciones descubiertas y las dadas por definición, es decir, relaciones de orden N (N-ORs) y relaciones paradigma (PR), una vez encontrado el nodo, se presenta el método de propagación hacia atrás (*backpropagation*) con el fin de modificar los valores sentimentales de nodos de orden superior, en base a la relevancia del tópico que estos contienen.

Se mostraron las contribuciones en cuanto a relevancia y componente sentimental a través del modelo respecto a variaciones de umbral de relevancia y ejecuciones del algoritmo, asimismo, se presenta una alternativa para perseguir las relaciones con más afectación sentimental en base a la relevancia.

Las ampliaciones al presente trabajo se centran en las siguientes tareas, (1) saciar las premisas asumidas de la existencia de los objetos tópico-sentimiento, (2) reportar resultados con distintas medidas de similitud semántica,(3) generar de manera automática los categorizadores, (4) optimizar la propagación de los componentes sentimentales.

Referencias

1. Lin, C., He, Y.: Joint Sentiment/Topic Model for Sentiment Analysis (2009)
2. Domingos, P., Pazzani, M.J.: On the optimality of the Simple Bayesian classifier under Zero-One loss, Machine Learning, 29 (2–3) (1997)
3. Lewis D.D.: Naive (Bayes) at forty: The independence assumption in information retrieval. In: Proc. of the European Conference on Machine Learning (ECML) (1998)
4. Blei, D.M. Ng, A.Y., Jordan, M.I.: Latent Dirichlet Allocation (2003)
5. Mei, Q., Ling, X., Wondra, M., Su, H., Zhai, C.X.: Topic Sentiment Mixture: Modeling Facets and Opinions in Weblogs (2007)
6. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up? Sentiment Classification using Machine Learning Techniques. In: Proceedings of EMNLP (2002)
7. Turney, P.D., Littman, M.L.: Unsupervised learning of semantic orientation from a hundred-billion-word corpus (2002)
8. Zagibalov, T., Carroll, J.: Automatic seed word selection for unsupervised sentiment classification of Chinese text. In: Proceedings of the 22nd International Conference on Computational Linguistics, Vol. 1 (2008)

9. Blei, D.M., Griffiths, T.L., Jordan, M.I.: The Nested Chinese Restaurant Process and Bayesian Nonparametric Inference of Topic Hierarchies (2010)
10. Kim, S., Zhang, J., Chen, Z., Oh, A., Liu, S.: A Hierarchical Aspect-Sentiment Model for Online Reviews (2013)
11. Rada M., Radev, D.: Graph-Based natural language processing and information retrieval (2011)
12. Wu, Z., Palmer, M.: Verb semantics and lexical selection. In: Proceedings of the 32nd Annual Meeting of the Associations for Computational Linguistics (1994)
13. Rada, R., Mili, H., Bicknell, E., Blettner, M.: Development and application of a metric on semantic nets. IEEE Transactions on Systems, Man, and Cybernetics 19 (1989)
14. Resnik, P.: Using information content to evaluate semantic similarity in a taxonomy. In: Proceedings of the 14th International Joint Conference on Artificial Intelligence, Vol. 1 (1995)
15. Blanchard, E., Harzallah, M., Briand, H., Kuntz, P.: A typology of ontology-based semantic measures (2005)