Método No Supervisado para la Sugerencia de Tags Utilizando Información Semántica Basada en Conocimiento

Alessandro Bokan Garay[1], Roque Lopez Condori [2]
alessandro.bokan@gmail.com, rlopezc27@gmail.com

[1]Universidad Católica San Pablo
[2]Universidad Nacional de San Agustín
Arequipa, Perú

Resumen: *En este artículo, se presenta un método no supervisado orientado a la sugerencia de tags para posts de blogs. El método propuesto tiene 3 etapas. En la primera, se crea la Base de Conocimiento con todos los post existentes del blog, en un intervalo de tiempo dado. En esta etapa, se aplica el proceso de etiquetado gramatical para extraer los sustantivos de todos los posts previamente seleccionados, y así poder generar una base de relaciones semánticas entre sustantivos y temas de cada post. En la segunda etapa, dado un post, se extraen los sustantivos y éstos son ponderados por frecuencia de aparición. En la última etapa, se hace un matching de los sustantivos ponderados con las relaciones semánticas ofrecidas por la Base de Conocimiento. Se aplica un algoritmo de ranking que otorga las palabras más importantes, las cuales serán sugeridas al autor del post como posibles tags. Para la comprobación de los resultados del algoritmo, se utilizó el método Gold Standard, en el cual una persona, experta en el tema, sugiere tags, y éstos se comparan con los tags sugeridos por el método no supervisado. Los resultados experimentales son satisfactorios, ya que el 70% de los tags sugeridos son efectivos para el autor.*

Abstract: *This paper presents an unsupervised method oriented to suggestion of tags for posts in blog. The proposed method has 3 stages. The first one creates the knowledge base with all existing blog post, in a given time interval. In this stage we apply the labeling process to extract grammatical nouns of all posts previously selected, and so, we generate a database of semantic relationships between nouns and topics associated with each post. In the second stage, given a post, we extract nouns and we send a list of words weighted by occurrence frequency. In the last stage, we do a weighted matching between nouns with the semantic relationships offered by the Knowledge Base. We apply a ranking algorithm that provides the most important words, which will be suggested to the post author. For checking the algorithm results, we used the Gold Standard method, in which a person, an expert on the topic, suggests tags, and these are compared with the tags suggested by the unsupervised method. The experimental results are satisfactory, because 70% of suggested tags are effective for the author.*

Palabras clave: *Procesamiento de Lenguaje Natural, sugerencia de tags, relaciones semánticas.*

1. Introducción

En la actualidad, la información en Internet está creciendo inmensamente. Un tema popular es el uso de *blogs*, o también llamados bitácoras digitales. Un blog es un sitio web que se encuentra en constante actualización, recopilando textos o artículos de distintos autores de forma cronológica. Al conjunto de documentos de un blog se denomina *posts*. Cada blog puede abarcar uno o varios temas, siguiéndose un esquema particular por cada uno, y cuyos autores poseen cierta libertad característica para plasmar sus pensamientos o comentarios. Dentro de un post existen temas relacionados con el contenido del mismo. Dichos temas son un conjunto de palabras denominados *tags*, y tienen como objetivo hacer referencia al texto escrito. Normalmente son insertados por los autores del post.

Debido al crecimiento de la información en Internet, el Procesamiento de Lenguaje Natural (PLN), un área de la Inteligencia Artificial (IA) que busca entender el lenguaje humano, ha tomado mucha importancia. Existen varias aplicaciones de PLN, tales como: recuperación de información, clasificación de texto [Rocchio, 1971], traducción automática [Paoineni et al, 2002], generación de resúmenes [Lin and Hovy, 2003], desambiguación del sentido de las palabras [Lesk, 1986; Schutze 1998], etiquetado automático (*tagging)*, entre otros.

La mayoría de las aplicaciones de PLN utilizan medidas de similitud para llevar a cabo sus tareas. Existen 2 tipos de medidas de similitud: léxica y semántica. La primera, realiza un matching produciendo un puntaje basado en la ocurrencia de unidades léxicas (palabras) entre dos textos. La similitud semántica, en cambio, es la similitud del significado de las palabras entre textos. Para esto es necesario basarse en una clase de información proporcionada por documentos de texto escritos con anterioridad, generándose así, una base de conocimiento con información precisa que ayudará a encontrar la mayor similitud semántica.

En el presente trabajo, se presenta un método no supervisado orientado a la sugerencia de tags. El algoritmo planteado fusiona los métodos de similitud léxica y semántica basándose en una Base de Conocimiento generada por todos los *tags* dentro de cada *post* (escrito con anterioridad).

El resto de este paper está organizado de la siguiente manera. En la sección 2, se describen los Trabajos Previos relacionados con nuestra propuesta. En la sección 3, se explica el Método Propuesto y los procesos que conforman cada una de sus etapas. En la sección 4, se presentan los Experimentos y Resultados obtenidos en el desarrollo del artículo. Las conclusiones del trabajo se encuentran en la sección 5. Por último, se listarán todas las referencias utilizadas en este paper.

2. Trabajos Previos

A continuación se presentan los trabajos con mejores resultados en la extracción de frases o *keywords (tags)* de un documento. Para esto, los trabajos aplicaron algoritmos de ranking basado en grafos, mostrando una mejora significativa cuando se hace uso de información semántica.

El *TextRank* es un algoritmo de ranking basado en grafos, derivado del algoritmo *PageRank*, utilizado por el buscador Google, cuya finalidad es extraer *keywords* [Mihalcea and Tarau, 2004]. Dado un texto, se seleccionan todas las palabras formando un grafo, donde: las palabras representan los vértices, las relaciones entre palabras de una misma oración representan las aristas. El peso de la arista es la frecuencia de coocurrencia de la relación en el texto. Una vez completado el grafo, se hace el ranking respectivo con la fórmula del *PageRank* [Brin and Page, 1998]. Para obtener los *keywords* se seleccionan las 'n' primeras palabras del ranking.

En el 2011, [Lopez and Tejada, 2011] crearon un método, denominado *MFSRank*, que busca seleccionar las *keyphrases* (frases clave) de un documento usando información semántica. El corpus semántico del cual se basaron se denomina ConceptNet. Al igual que el método BUAPx propuesto por [Ortiz et all, 2010], el algoritmo MFSRank selecciona todas las Secuencias Maximales Frecuentes (SMF) [Agrawal and Srikant, 1995]. Luego, se otorga un peso a la relación entre dos SFM ubicados en una misma oración. La propuesta de este algoritmo es agregar un valor semántico otorgado por las relaciones semánticas que ofrece Concepnet. Por último, realiza el ranking aplicando el algoritmo PageRank.

3. Método Propuesto

El método propuesto tiene 3 etapas. En la primera, se crea la Base de Conocimiento seleccionando todos los *posts* de un blog para un intervalo de tiempo dado. Aquí se aplica el proceso de etiquetado gramatical para la extracción de sustantivos de todos los *posts* seleccionados. Se extraen los sustantivos para generar una base de relaciones semánticas entre sustantivos y temas asociados de cada *post*. En la segunda etapa, se extraen los sustantivos del texto y se ponderan por frecuencia de aparición. En la tercera etapa, se hace un *matching* entre los sustantivos ponderados y las relaciones semánticas ofrecidas por la Base de Conocimiento, aquí se aplica un algoritmo de ranking para obtener los temas más importante, los cuales serán sugeridos al autor del *post*.

Cabe resaltar que, tanto para la primera etapa, como para la segunda, se ha realizado un pre procesamiento de los documentos (eliminación de *stopwords* y signos de puntuación). Para visualizar el método propuesto se creó una figura nombrando los pasos a seguir, observar la Figura 1.

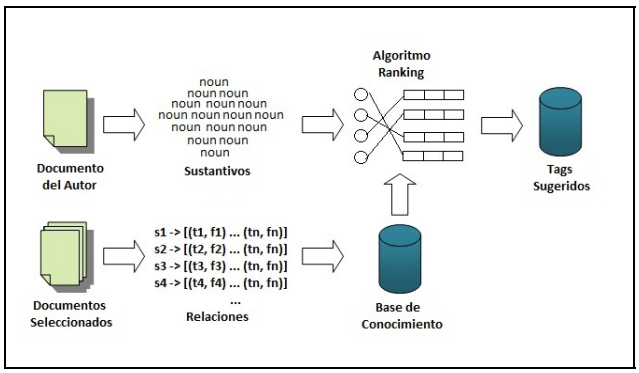

Figura 1. Método Propuesto.

3.1. Pre procesamiento

El objetivo de este proceso es eliminar datos de los *posts* que no son relevantes, es decir, que no aporten significado alguno para la solución del problema. Esta etapa consta de los siguientes pasos:

- Eliminación de Stopwords

Consiste en la eliminación de palabras que no aporten información, ya que, por su naturaleza, no son relevantes al momento de sugerir palabras en un determinado contexto. Por ejemplo: conjunciones, pronombres, artículos, preposiciones, etc. Estas palabras suelen aparecer frecuentemente en los textos. Se considera que una palabra que aparece en al menos un 80% de todos los documentos de una colección en particular, es inútil para el proceso de recuperación de información.

- Eliminación de signos de puntuación

Los signos de puntuación son delimitadores entre frases o párrafos, que establecen una jerarquía sintáctica de proposiciones, logran estructurar el texto. El objetivo de eliminar símbolos de puntuación es limpiar el texto, y así, obtener solamente las palabras.

3.2. Selección de términos (Part of Speech Tagging)

El etiquetado gramatical es un proceso de la Lingüística Computacional que consiste en asignar o etiquetar a cada una de las palabras de un texto por su categoría gramatical. Una categoría gramatical se define por el comportamiento sintáctico o morfológico del elemento léxico en función del contexto en que aparece, es decir, una palabra puede poseer una o distintas categorías gramaticales dependiendo del contexto que se encuentre en el texto. Sin embargo, debido al pre procesamiento realizado con anterioridad, resulta simple realizar un etiquetado gramatical de un conjunto de palabras para, de esta manera, seleccionar sólo los *sustantivos* del texto.

3.3. Primera Etapa: Base de Conocimiento

Como se mencionó anteriormente, un blog es un sitio web que recopila textos de forma cronológica y que a su vez está en constante actualización. La idea de este trabajo es crear una Base de Conocimiento (BC) formada por *relaciones semánticas* entre palabras (sustantivos) y temas (*tags)*, en un intervalo de tiempo dado. Este intervalo de tiempo varía según discontinuidad de temas que se va dando en la vida de un blog, por ejemplo: un tema de

política podría ser discutido y controversial durante un mes y medio aproximadamente. La BC deberá *volver a crearse* en cada intervalo de tiempo dado.

A medida que se va creando la BC, se va realizando el *pre procesamiento* de la información proporcionada por cada blog dentro del conjunto de posts seleccionados para luego pasar al etiquetado gramatical de cada texto. Nuestra hipótesis afirma que: "*al seleccionar sólo los sustantivos de cada post, se generará un menor número de relaciones semánticas, pero que están asociadas a unidades léxicas significativas (sustantivos) que podrían representar el contexto del post.* El trabajo plantea que un conjunto de sustantivos puede representar el tema del que está refiriendo el texto.

Para formar la base de conocimientos, se seleccionan todos los posts de una *categoría* (política, deporte, actualidad, economía, etc.). Una vez seleccionados los *posts*, se realiza un algoritmo que genera relaciones entre una palabra (sustantivo) y el conjunto de tuplas *(ti, fi)*, donde *ti* es el *tag* asociado al post (donde está ubicada la palabra) y *fi* es la frecuencia de repetición del *tag* en todo el conjunto de posts.

Dado un conjunto de post seleccionados de una categoría determinada y en un intervalo de tiempo, se realiza el siguiente algoritmo:

Algoritmo 1 Creación de la Base de Conocimiento

Requiere: Conjunto de post selecionados / Base de Conocimiento (BC)

```
para cada post en posts_selecionados hacer
    contenido <= post -> contenido
    contenido <=  eliminar_stopwords(contenido)
    contenido <= eliminar_signos(contenido)
    sustantivos<= obtener_sustantivos(contenido)
    para cada sustantivo en sustantivos hacer
        si sustantivo está en BC entonces
            para cada tag en post -> tags hacer
                si tag está relacionado al sustantivo entonces
                    incrementar _frecuencia(sustantivo, tag)
                sino
                    asignar_relacion(sustantivo, tag, freq <= 1)
                fin si
            fin para
```

3.4. Segunda Etapa: Ponderado de Término

En esta etapa, de igual manera, se realiza un pre procesamiento de texto. Luego se seleccionan los sustantivos, los cuales serán ponderados de acuerdo con su frecuencia de aparición en el texto.

Para asignar un valor significativo a cada término, es necesario basarse de una técnica de ponderación, en este caso, el Término Frecuente (TF). Este ponderado toma en cuenta que un término que ocurre con frecuencia en un documento puede reflejar mejor el contenido del documento que un término que se produce con menos frecuencia [Luhn, 1957]. La ponderación *TF* consiste en evaluar el número de veces que la palabra aparece en el documento y asignar un mayor peso a los términos con mayor frecuencia.

$$d_i(t_j) = f_{ij}$$

Donde *fi,j* es el término frecuente *j* en el documento *i*

3.5. Tercera Etapa: Algoritmo Ranking

Una vez obtenido el conjunto de sustantivos ponderados, se hace un *matching* con las relaciones semánticas de la Base de Conocimiento. El algoritmo asigna a cada sustantivo del texto sus respectivos *tags* asociados (a través de la relación semántica) multiplicando la frecuencia de aparición del sustantivo con la frecuencia de aparición del *tag*. Así se obtienen una lista de *tags* ponderados, brindados por la relación, que llamaremos *sugerencias*. Si en caso algún sustantivo no posea *tags* relacionados, se agregará dicho *sustantivo* a la lista de tags sugeridos (*sugerencias*). A continuación, el algoritmo de ranking.

Algoritmo 1 Creación de la Base de Conocimiento

Requiere: Conjunto de post selecionados / Base de Conocimiento (BC)

```
para cada post en posts_selecionados hacer
    contenido <= post -> contenido
    contenido <=  eliminar_stopwords(contenido)
    contenido <= eliminar_signos(contenido)
    sustantivos<= obtener_sustantivos(contenido)
    para cada sustantivo en sustantivos hacer
        si sustantivo está en BC entonces
            para cada tag en post -> tags hacer
                si tag está relacionado al sustantivo entonces
                    incrementar _frecuencia(sustantivo, tag)
                sino
                    asignar_relacion(sustantivo, tag, freq <= 1)
                fin si
            fin para
```

Al final, se obtiene una lista de tags sugeridos (con elementos repetidos) con un ranking respectivo. Para obtener un conjunto, se debe hacer un promedio de los rankings de los elementos repetidos. Por último, se ordenan los elementos por el mayor ranking. De esta manera, se obtiene un conjunto de *tags* que serán sugeridos al autor del post.

4. Experimentos y Resultados

4.1. Datos de Prueba

Para la creación de la Base de Conocimiento, se recopilaron *posts* de un blog de noticias con las siguientes categorías: política, internacional, economía y tecnología. Se recogió un total de 200 posts, es decir, 50 posts por categoría, cada uno con su respectiva lista de *tags* asociados. Se obtuvo un total de 637 *tags* distintos. El intervalo de tiempo utilizado fue de 2 meses.

Se escogieron 2 autores, cada uno escribió un *post* por categoría con una lista de 10 *tags* asignados por cada uno.

El algoritmo retorna una lista de 15 *tags* sugeridos. Aplicando el método *Gold Standard*, el autor propone 3 niveles de comparación: *I* (iguales a los tags insertados), *A* (aceptables), y *M* (malos o fuera de contexto).

4.2. Resultados

Los resultados obtenidos son satisfactorios. Ver Tabla 1.

Tabla 1. Resultados del algoritmo.

	Política			**Internacional**			**Economía**			**Tecnología**		
Nivel	*I*	*A*	*M*	*I*	*A*	*M*	*I*	*A*	*M*	*I*	*A*	*M*
Autor 1	5	6	4	5	3	7	6	6	3	5	5	5
Autor 2	5	4	6	4	5	6	6	7	2	4	5	6

Para el autor 1, aproximadamente un 35% de los tags sugeridos son iguales a los tags insertados por el autor, el 33,3% son aceptables y el 31,7% son considerados como malos. Para el autor 2, aproximadamente un 31,7% son iguales, 35% son aceptables y 33,3% son malos. Haciendo un promedio, se obtiene un total de 33,35% de igualdad, 34,15% de aceptación y 32,5% de error. En conclusión, si sumamos el porcentaje de igualdad con el de aceptación, obtenemos aproximadamente un 70% de efectividad del algoritmo.

Cabe resaltar que el criterio de las persona es variable, sin embargo, el algoritmo ha demostrado hacer una sugerencia coherente al post escrito.

5. Conclusiones

En este artículo, se ha presentado un método no supervisado orientado a la sugerencia de *tags* para posts de blogs. La novedad de este trabajo radica en la utilización de una Base de Conocimiento (BC).

Es importante resaltar que el algoritmo obtiene buenos resultados al formar una BC por categorías diferentes, ya que las relaciones reflejan un contexto diferente por cada categoría, y el uso de sustantivos específica a las relaciones

5.1. Trabajos futuros

Entre los principales trabajos futuros, nosotros encontramos: (1) Utilizar una base de conocimiento de dominio específico. (2) Utilizar otros mecanismos de pesado de términos, tales como IDF y TF-IDF.

Referencias Bibliográficas

[Agrawal and Srikant, 1995] Agrawal, R. and Srikant, R. (1995). Mining sequential patterns. Proceedings of the Eleventh International Conference on Data Engineering, 27:1-22.

[Brin and Page1998] S. Brin and L. Page. 1998. The anatomy of a large-scale hypertextual web search engine. In *COMPUTER NETWORKS AND ISDN SYSTEMS*, 107–117. Elsevier Science Publishers B. V.

[Chong et all, 2006] Chong, H., Yonghong, T.,Zhi, Z.,Charles X. L., Tiejun, H. 2006. Keyphrase extraction using semantic networks structure analysis. In Proc. of the ICDM06. 275–284.

[Ian et all, 1999] Lan, H. W., Gordon W. P., Eibe F., Carl G., and Craig G. 1999. KEA: practical automatic keyphrase extraction. In Proceedings of the fourth ACM conference on Digital libraries (DL '99). ACM, New York, NY, USA, 254-255.

[Lesk, 1986] Lesk, M. 1986. Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. In Proceedings of the SIGDOC Conference 1986.

[Lin and Hovy, 2003] Lin, C., and Hovy, E. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In Proceedings of Human Language Technology Conference.

[Lopez and Tejada, 2011] Lopez, R. and Tejada, J. MFSRank: An unsupervised method to extract keyphrases using semantic information. Mexican International Conference on Artificial Intelligence, 1-7.

[Luhn, 1957] Luhn, H. P. 1957. A statistical approach to mechanical encoding and searching of literary information. IBM Journal of Research and Development, 1(4):309-317.

[Mihalcea and Tarau, 2004] R. Mihalcea and P. Tarau. 2004. TextRank: Bringing order into texts. In *EMNLP 2004, ACL*, 404–411.

[Ortiz et all, 2010] Ortiz R, Pinto D, Tovar M, Jiménez H. BUAP: An unsupervised approach to automatic keyphrase extraction from scientific articles. In Proceedings of the 5th International Workshop on Semantic Evaluation (SemEval '10). Association for Computational Linguistics, Stroudsburg, PA, USA, 174-177.

[Papineni et all, 2002] Papineni K, Roukos S, Ward T, and Zhu, W. 2002. Bleu: a method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics.

[Rocchio, 1971] Rocchio J. (1971). Relevance feedback in information retrieval. Prentice Hall, Ing. Englewood Cliffs, New Jersey.

[Schutze, 1998] Schutze, H. (1998). Automatic word sense discrimination. Computational Linguistics 24(1):97-124.