Evaluando Chi-Cuadrada y PageRank en la Extracción de Bigramas Clave desde Texto Plano

Franco Rojas-López, Jorge Jaime Juárez Lucero, José Luís Hugo Díaz, y María del Rayo Graciela Guevara Villa

franco.rojas, jorge.jaime, hugo.diaz rayo.guevara@metropoli.edu.mx

Resumen

La extracción de términos clave desde texto plano sigue siendo una tarea de investigación abierta para el éxito de diferentes aplicaciones del Procesamiento del Lenguaje Natural. En este artículo proponemos una técnica para la extracción de bigramas clave desde texto plano en el idioma español usando un enfoque basado en grafos. La técnica implementa la chi-cuadrada para ponderar bigramas los cuales son usados para asignar un peso a la arista que une nodos en un grafo no dirigido. Nosotros evaluamos la técnica propuesta en documentos con un promedio de 757 palabras. Los resultados obtenidos son alentadores porque por inspección los bigramas recuperados describen muy bien el tema del cual trata el documento.

Palabras clave: Resumen, chi-cuadrada, grafo, PageRank

1. Introducción

Las palabras clave comúnmente son usadas por los motores de búsqueda (Vibhanshu & Kartik, 2007), Desambiguación del Sentido de la Palabra (Rojas López, López Arévalo, & Sosa Sosa, 2012), Clasificación de texto entre otras (Phayung, Pudsadee, & Vatinee, 2011) entre otras. Leer y resumir el contenido de grandes cantidades de texto en un conjunto pequeño de temas es una tarea difícil y requiere de un gran esfuerzo humano, de tal modo que llega ser imposible con la creciente cantidad de información disponible en formato digital. Como resultado de la creciente información, los sistemas automatizados están siendo usados para generar resúmenes, o encontrar los temas de los que trata un documento. Esta tarea es compleja y desafiante debido a las propiedades intrínsecas del lenguaje natural, así como la dificultad inherente en determinar si una palabra o conjunto de palabras representan de manera precisa los temas presentes en el documento.

En el trabajo propuesto combinamos la técnica chi-cuadrada y un enfoque basado en grafos, para recuperar bigramas clave, como una tarea intermedia para extraer un resumen desde texto plano. La tarea es desafiante porque la técnica es aplicada en documentos cortos de un tamaño de 5 KiB y en promedio 800 palabras.

El documento está estructurado de la siguiente manera. En la sección 2 se presenta una breve descripción de investigaciones relacionadas con el trabajo propuesto. La sección 3 describe la metodología propuesta. En la sección 4 se muestran los resultados obtenidos de la metodología propuesta. Finalmente las conclusiones y trabajo futuro se presentan en la sección 5.

2. Estado del Arte

El rápido crecimiento en las tecnologías de la información han favorecido la publicación de cantidades inmensas de información digital, tan solo Google tiene indexadas cerca de 47 billones de páginas web1. Por lo tanto el crecimiento de la información exige nuevos algoritmos que permitan ayudar al usuario en la búsqueda de información así como comprender el tema contenido en el documento de su interés. (Rafeeq, 2010) propuso una técnica para la extracción de resumen desde texto plano. La técnica consiste de 4 fases, en la primera fase se lleva a cabo la tarea de pre-procesamiento del texto (remover stopwords tales como: conjunciones, pre proposiciones, artículos, etc.). La segunda fase consiste en recuperar frases claves. En la tercera fase se extraen las sentencias más importantes y la cuarta fase reduce las sentencias extraídas en la fase anterior para generar el resumen. Márius et al. (Ajgalik, Barla, & Bieliková, 2013) propusieron un algoritmo para la extracción de conceptos clave usando las palabras más representativas del texto ellos representan vectores de conceptos en lugar de vectores de palabras usando información mutua. Otro trabajo importante en la extracción de resúmenes desde texto plano fue propuesto por Rada y Paul (Mihalcea & Tarau, 2004). Ellos asignan una ponderación a los términos del documento usando TF-IDF. Esta ponderación es usada para construir un grafo no dirigido con pesos en las aristas. Finalmente ejecutan el algoritmo TextRank (Mihalcea & Tarau, 2004) para encontrar los términos clave y finalmente extraer un resumen desde texto plano. El trabajo que se propone es similar al reportado por Rada y Tarau, la diferencia radica en que nosotros modificamos e implementamos la chi-cuadrada para ponderar bigramas. El grafo es construido usando los

[1] http://www.worldwidewebsize.com/, visitado en diciembre de 2015

bigramas que cumplen la condición de transitividad. La metodología propuesta se explica en siguiente sección.

3. Metodología

La extracción de términos clave, es considerada una tarea intermedia para el éxito de diversas tareas del Procesamiento del Lenguaje Natural tales como en Recuperación de Información, Desambiguación del Sentido de la Palabra, Generación de Resúmenes, etc.
En este artículo se implementa una técnica para la extracción de términos clave desde texto plano. Tal técnica consiste en medir el grado de correlación entre un par de palabras para encontrar bigramas en un documento. Para lograr este objetivo se implementó la técnica x^2 y PageRank, las cuales han sido usadas por su efectividad en diferentes tareas del PLN (Rojas López, López Arévalo, & Sosa Sosa, 2012, (Mihalcea & Tarau, 2004)).

3.1 Chi-Cuadrada

Una alternativa para medir la independencia entre dos variables es la Chi-cuadrada la cual no asume probabilidades distribuidas normalmente. Cabe mencionar que no es de nuestro interés discutir los aspectos estadísticos de la x2 una explicación más extensa puede verse en el libro de Manning y Schütze (D. Manning & Schutze, 1999). El valor x2 entre dos términos se obtiene empleando la ecuación 1.

$$X^2 = \frac{(O - E)^2}{E} \qquad (1)$$

Donde O denota la frecuencia observada y E denota la frecuencia esperada entre los términos w_i y w_{i+1}. La frecuencia esperada se obtiene aplicando la Ecuación 2. Donde $f(w_i;w_{i+1})$ es la frecuencia de aparición del bigrama w_i y w_{i+1} en el documento usando una ventana de 2 términos; es decir, dos términos después de la palabra de interés. La probabilidad de ocurrencia de los términos w_i y w_{i+1} se obtiene dividiendo su valor de frecuencia entre el número de ocurrencia de todos los pares de bigramas encontrados en el documento.

$$E = \frac{f(w_i, w_{i+1})}{\sum t} \qquad (2)$$

La frecuencia observada de los términos w_i y w_{i+1} se obtiene multiplicando los factores *a* y *b* de la Ecuación 3. $f(w_i, x)$ y $f(x, w_{i+1})$ es la frecuencia del término w_i y

w_{i+1} con un término *x* cualquiera en el documento de texto. $\sum f(w_i)$ y $\sum f(w_{i+1})$ es la frecuencia de ambos términos en el documento de texto.

$$a = \frac{f(w,x)}{\sum f(w_i)}, \qquad b = \frac{f(x, w_{i+1})}{\sum f(w_{i+1})} \tag{3}$$

3.2 Construcción del grafo

En la literatura se han reportado diferentes algoritmos para encontrar el nodo más importante en un grafo, tales como: *Indegre*, el cual determina la importancia de un nodo considerando los enlaces que convergen en él. *Key Problem Player* aquí un vértice en importante si es relativamente cercano al resto de los nodos en el grafo (Navigli & Lapata, 2007), entre otros. El procedimiento para construir el grafo se describe a continuación. Sea $C_b = \{(w_1, w_2), (w_1, w_3), \ldots, (w_2, w_3), \ldots, (w_i, w_{i+1})\}$ el conjunto de bigramas extraídas desde el texto plano. Estos bigramas constituyen los nodos en el grafo. La representación del grafo está dada por $G = (V; E; W)$, donde *V* son los vértices (bigramas), *E* son las aristas (relaciones entre vértices) y *W* es un valor numérico que asocia dos bigramas.

3.3 PageRank personalizado

PageRank Personalizado (Personalized PageRank, (PPRank)) (Agirre & Soroa, 2009, (Mihalcea & Tarau, 2004)), es un medida basada en grafo modificada de la versión original PageRank propuesto por Brin y Page (Sergey & Lawrence, 107-117), la cual consiste en un grafo no dirigido con pesos en las aristas. Después de ejecutar el algoritmo, un score es asociado con cada vértice como se muestra en la Ecuación 4.

$$PPRank(v_i) = (1-\alpha) + \alpha * \sum_{vj\ \varepsilon\ In(vi)} (w_{ij} / \sum_{VK\ \varepsilon\ Out(VJ)} w_n(k)) PPRank(v_j) \tag{4}$$

De acuerdo a la literatura el valor del factor α es de 0.85, el cual es usado en esta evaluación. w_{ij}, es el peso entre el vértice *j* e *i*. $In(v_i)$ y $Out(v_j)$ son el grado de entrada y salida de los vértices v_i y v_j.

4. Experimentos y resultados

Esta sección presenta los resultados de evaluación de la Metodología propuesta. Es importante mencionar que por el momento no contamos con un *dataset* de términos clave en español para validar el algoritmo. Por lo tanto en las pruebas usamos documentos del dominio de la naturaleza, particularmente los resultados

que se muestran son de un documento cuyo tamaño es de 5 KiB con un total de 757 palabras. Antes de ejecutar el algoritmo los documentos fueron divididos en sentencias usando la herramienta CoreNLP[2]. A continuación las palabras denominadas stopwords[3] fueron eliminadas del documento. Cabe mencionar que no usamos un etiquetador para identificar la categoría gramatical de las palabras con el objetivo de hacer más eficiente el algoritmo aunque se pretende integrarlo a futuro.

4.1 Chi-Cuadrada

La Tabla 1, muestra el resultado obtenido después de aplicar el procedimiento descrito en la subsección 3.1. Los resultados que se muestran son sorprendentemente buenos solo el bigrama naturaleza-reside parece no tener coherencia para ser aceptado como bigrama. Aunado a lo anterior los bigramas ponderados describen muy bien el tema del documento lo cual es sumamente importante para extraer un resumen o identificar el tema del cual trata el documento.

Bigrama	Valor x^2
Medio-ambiente	0.015122
Naturaleza-reside	0.011834
Cuidar-vegetación	0.009862
Cuidar-fauna	0.009862
Protección-naturaleza	0.009862
…	…

Tabla 1. *Bigramas obtenidos usando x^2*

4.2 Personalizado PageRank

Para aplicar el algoritmo PPRank un grafo fue construido como fue descrito en la subsección 3.2. El objetivo de aplicar tal algoritmo fue dar mayor soporte a los bigramas recuperados. Después de recuperar los bigramas desde texto plano C_b (subsección 3.2), cada bigrama fue buscado en la lista que se muestra en la Tabla 1. Si el bigrama no existe en el grafo entonces el bigrama es agregado como

[2] http://stanfordnlp.github.io/CoreNLP/, visitada en diciembre de 2015

[3] Las stopwords, son palabras de significado vacío como los artículos, los pronombres o las preposiciones.

vértice. Supongamos que primero se agrega como vértice el bigrama (a-b) para insertar otro vértice este debe cumplir la propiedad de transitividad; es decir, debe ser un bigrama de la forma (b-c). Creemos que esto es importante porque es muy probable que el término *a* también este asociado con el término *c*. Una vez insertados ambos vértices se traza una arista para unirlos. El peso asignado en la arista se obtiene de la Tabla 1, para ello se suma el valor de x^2 de ambos bigramas y se divide por 2. La Tabla 2 muestra los resultados obtenidos después de ejecutar el algoritmo PPRank.

Bigrama	Valor PPRank
paisajística-protección	0.13101
una-protección	0.12883
económicamente-protección	0.08099
puesto-protección	0.08054
protección-especies	0.07735
…	…

Tabla 2. *Bigramas obtenidos usando PPRank*

Los resultados obtenidos no muestran una mejora al ejecutar el algoritmo PPRank. Por inspección podemos notar que los bigramas recuperados no describen el tópico del documento como lo hace la técnica x^2, por lo que es importante buscar una mejora para el algoritmo PPRank, cambiando la forma en la cual se construye el grafo.

5. Conclusiones y trabajo futuro

En este trabajo se propuso un enfoque para la extracción de bigramas clave desde texto plano para el idioma español, usando x^2 y el algoritmo PPRank. Los mejores resultados se obtuvieron al evaluar x^2, por lo que se propone como trabajo futuro proponer una alternativa para construir el grafo y en consecuencia mejorar los resultados obtenidos. Aunado a lo anterior se propone desarrollar una aplicación para extraer un resumen desde el texto plano, tal aplicación será integrada en un Software el cual genera una revista digital a partir de texto. El Software que genera la revista así como la aplicación que extrae el resumen estará disponible en línea para que el usuario tenga conocimiento del tema de la revista sin necesidad de leer el documento completo.

Referencias

Agirre, E., & Soroa, S. (2009). Personalizing pagerank for word sense disambiguation. *In Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, 33-41.

Ajgalik, M., Barla, M., & Bieliková, M. (2013). From ambiguous words to key-concept extraction. . *In Database and Expert Systems Applications (DEXA), 2013 24th International Workshop on*, 63-67.

D. Manning, C., & Schutze, H. (1999). *Foundations of Statistical Natural Language Processing.* MIT Press, Cambridge, MA, USA.

Mihalcea, R., & Tarau, P. (2004). TextRank: Bringing Order into Texts. *In Proceedings of Empirical Methods for Natural Language Processing*, 404-411.

Navigli, R., & Lapata, M. (2007). Graph connectivity measures for unsupervised word sense disambiguation. *In Proceedings of the 20th International Joint Conference on Artificial Intelligence*, 1683-1688.

Phayung, M., Pudsadee, B., & Vatinee, N. (2011). A chisquare-test for word importance differentiaton in text classification. *In Proceedings of International Conference on Information and Electronics Engineering*, 110-114.

Rafeeq, A.-H. (2010). Text Summarization Extraction System (TSES) Using Extracted Keywords. *International Arab Journal of e-Technology* , 164-168.

Rojas López, F., López Arévalo, I., & Sosa Sosa, V. (2012). Combining local and related context for word sense disambiguation on specific domains. *In International Conference on Data Management Technologies and Applications*, 135-140.

Sergey, B., & Lawrence, P. (107-117). The anatomy of a large-scale hypertextual web search engine. . *Computer Networks and ISDN Systems*, 1998.

Vibhanshu, A., & Kartik, H. (2007). Keyword generation for search engine advertising using semantic similarity between terms. *In Proceedings of the ninth international conference on Electronic commerce* , 89-94.