Uso de Grafos con Semántica para Minería de Texto

María de Jesús Estudillo Ayala

Academia de Computación, Centro de Bachillerato Tecnológico Industrial y de Servicios No. 178, chuyest@hotmail.com

1.Resumen

Presentamos un método que busca mejorar la eficacia de los sistemas de minería de texto que utilizan aprendizaje de conceptos basados en grafos. Se incorpora mecanismos semánticos a la búsqueda de patrones en texto (en inglés). Los resultados que se presentan se analizan en forma cualitativa y cuantitativa, la primera mediante la observación de los grafos obtenidos por cada experimento y la segunda mediante clasificación de texto.

2.Introducción

El conocimiento o información para el ser humano es fundamental. Existen en el mundo volúmenes inmensos de información expresada en lenguaje natural (inglés, español, latín, etc.), almacenada en diferentes medios. Para obtener este conocimiento nos enfrentamos al problema de la administración de información, es decir, buscarla, seleccionarla y utilizarla de la mejor forma y en el menor tiempo posible[9].

La lingüística computacional, en particular el procesamiento automático de textos junto con la minería de datos para colecciones grandes de texto, se enfoca en la solución de este gran problema, dando como resultado lo que se conoce como **minería de texto**. La minería de texto se encarga de buscar patrones en texto de lenguaje natural, y se puede definir como el proceso de analizar el texto para extraer información para propósitos particulares[2] y [5].

En este artículo se propone una metodología para el descubrimiento de patrones en la minería de texto, la cual se basa en el descubrimiento del conocimiento en bases de datos[8], pero en éste los datos son texto y se incorporan técnicas de lenguaje natural, como es el análisis gramatical e incorporación de semántica al texto. Para obtener los patrones se utilizó el algoritmo de SubdueCL, para el análisis gramatical el programa TreeTagger y para la semántica WordNet, los cuales se describen en este artículo.

3.SubdueCL

SubdueCL es un sistema de aprendizaje de conceptos, tiene la capacidad de aprender conceptos a partir de ejemplos "positivos" y "negativos" y obtiene como resultado subestructuras que describen los ejemplos positivos pero no a los negativos [8].

4.WordNet

WordNet, es la base de datos léxica mejor desarrollada y extensamente usada para el inglés[1].

Está compuesta por tres bases de datos, en una están los sustantivos, en otra los verbos, y en una tercera los adjetivos y los adverbios; estas se organizan en sinónimos. Por medio de diversas relaciones se unen a los sinónimos [4].

Las relaciones que se encuentran en WordNet son: sinonimia, hiperonimia, hiponimia, meronimias, elemento de colectividad, sustancia, causa, implicación, derivación, similitud[6]. En este trabajo hacemos uso de la relación de sinonimia y de la hiperonimia, donde la primera es el conjunto de palabras que en un contexto dado expresan un concepto y la segunda son relaciones de clases a subclases[1].

5.TreeTagger

TreeTagger es una herramienta que se utiliza para marcar los textos con etiqueta en diferentes lenguajes como son el francés, alemán, inglés, italiano y griego, es fácilmente adaptable a otras idiomas[7]. La notación que se utiliza para las etiquetas es tomado del proyecto Penn Treebank como se muestra en la Tabla 1. El proyecto de Penn Treebank le agrega al texto natural una estructura lingüística, dependiendo de su gramática[7.

Word	Pos	Lemma
The	DT	The
TreeTagger	NP	TreeTagger
Is	VBZ	Be
Easy	JJ	Easy
To	TO	To
Use	VB	Use
	SENT	.

Tabla 1: Ejemplo del etiquetamiento

6.Método para el descubrimiento de patrones de texto

El método que se utiliza para el descubrimiento de patrones en la minería de texto se muestra la figura 1 y se describe a continuación:

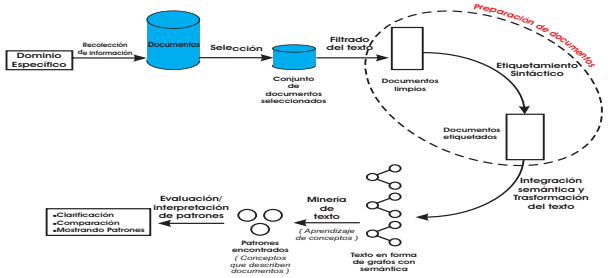


Figura1: Método para el descubrimiento de patrones de texto

El método funciona de la siguiente manera: Primero, en la *Recopilación de Información* se buscan documentos relativos a un domino específico sobre el que se desean encontrar los patrones, por ejemplo documentos de animales, computación, contabilidad, arquitectura, medicina, etc. Entonces, estos documentos se Seleccionan y se agrupan en base al dominio que representan por medio de la asignación de una clase a cada documento, esta parte de nuestro método es llamada *Selección*. A los documentos previamente seleccionados y agrupados se les eliminan elementos referentes al formato del archivo a procesar, como puede ser, el encabezado, etiquetas de html o latex, caracteres especiales de formato (\$, \& , \% , crtl M , etc.) a este paso se le conoce como *Filtrado del texto*. Enseguida, se etiquetan las palabras dependiendo de su papel sintáctico dentro de cada oración. Al filtrado y etiqueta miento de los documentos lo llamamos *Preparación de los documentos*. A continuación se toman todas las palabras que han sido etiquetadas sintácticamente tales

como sustantivos y se les buscan sus sinónimos o hiperonimia se crea un grafo por documento con estas palabras, es decir, los sustantivos junto con sus sinónimos o hiperónimos se transforman a grafo, en esto consiste la *Integración semántica y transformación del texto*. Estos grafos alimentan a los algoritmos de aprendizaje de conceptos basados en grafos (Subdue y SubdueCL), para obtener los patrones, es decir se realiza el proceso de *minería de texto* . Por último se *Evalúan e interpretan los patrones*, esto es, se hace un análisis de los patrones obtenidos comparándolos entre los resultados obtenidos en diferentes pruebas, y también se evalúan los patrones mediante clasificación de texto con la técnica de validación cruzada de 10 pasos (10 fold cross validation),

7.Resultados experimentales

Para los experimentos se utilizaron dos Bases de Datos, una con información sobre resúmenes de artículos, los cuales fueron divididos en artículos de computación (positivos) y artículos de cualquier otra área del conocimiento(negativos). La segunda base de datos usada contiene información sobre la descripción de animales y esta fue dividida en descripción de animales mamíferos (positivos) y animales no mamíferos(negativos): reptiles, aves, peces. Las bases de datos se obtuvieron manualmente buscando y seleccionando la información en internet. La distribución de los datos se muestra en la tabla 2.

Documentos	Cantidad de archivos
Resúmenes de computación(RC)	25
Resúmenes de no computación(RNC)	25
Descripción de Mamíferos	134
Descripción de no mamíferos	82

Tabla2: Distribución de los documentos

Se realizaron cinco tipos de experimentos los cuales solo se muestra el primero y el cuarto

- En el primer experimento se obtienen los grafos estrella utilizando todas las palabras en los documentos. En la figura 2 se muestran las tres mejores subestructuras obtenidas en este experimento.

Figura 2: Grafos obtenidos en el primer experimento.

- En este cuarto experimento se emplea el método propuesto, en el cual, se extraen por palabra un sinónimo por cada una de las distintas definiciones extraídas de WordNet. En la figura 3 se muestran las tres mejores subestructuras obtenidas en este experimento.

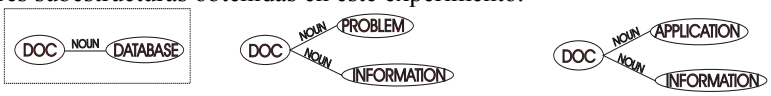
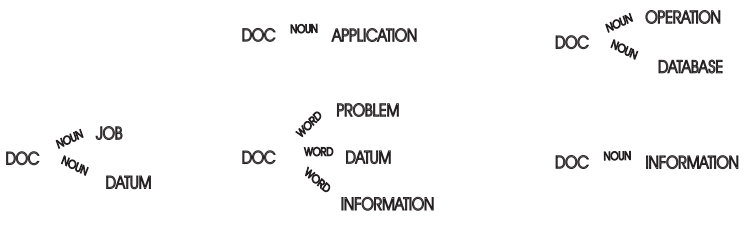


Figura 3: Grafos obtenidos en el cuarto experimento

Observe la diferencia entre el primer experimento al cuarto, en el primer experimento las subestructuras muestran palabras que son muy repetitivas en cualquier documento, por lo que, estos grafos no describen bien el contenido de los documentos. Este problema no se tiene con el cuarto experimento y la cantidad de grafos aumenta al triple de los encontrados en el primer experimento.

La clasificación de texto, se realiza con los grafos obtenidos en los experimentos anteriores. Se realizó una validación cruzada de 10 pliegues (10-fold cross-validation por cada prueba. Los resultados obtenidos se muestran en la tabla 3.



Prueba	RC vs RNC	DM vs DNM
1ra.	68%	-
2da.	70%	-
3ra.	69.15%	76.66%
4ta.	69.15%	75.05%
5ta.	67.5%	75.26%

8.Conclusiones y trabajos futuros

La obtención de patrones utilizando el método propuesto en el que se incorpora semántica produce más y mejores grafos que describen al texto, sin embargo al utilizar estos patrones para realizar clasificación de texto el porcentaje de eficiencia no mejora, disminuye el 1.5%. La disminución en la clasificación puede deberse a que al agregar información que ayuda a la tarea de obtención de patrones (sinonimia e hiperonimia) se está agregando información que no ayuda a la tarea de clasificación. Como trabajo futuro queda la tarea de hacer un análisis más profundo sobre calidad de patrones y clasificación; realizar pruebas con otras bases de datos, utilizar estructuras de grafos con más información gramátical.

Referencias

[1] C. Fellbaum. WordNet: An Electronic Lexical Database. 1998.

[2] Hearst. Untangling text data mining. Pro. Of ACL'99: Tje 37th Annual Meeting of the Association for Computational Linguistics, University of Maryland 1999.

[3] Lawrence B. Holder, Jesus A. González and Diane J. Cook. Graph-based concept learning. Proceding of the Fourteenth Anual Florida AF Research Symposium, 2001.

[4] Daniel Jurafsky and James H. Martin. Speech and Language Processing. 2000.

[5] Kodratff. Knowledge discovery in texts: A definition and applications. Proc. Of the 11th International Symposium of Fundations of Intelligent Systems (ISMS-99), 1999.

[6] George A Miller. Wordnet: a lexical database for the anglis language, http://www.cogsci.princeton.edu/wn/index.shtm , 2002

[7] Helmut Schmid. Treetagger a lenguage independent part-of-speecht tagger. http://www.ims.unistuttgart.de/projekte/corplex/ , 2002

[8] Pashraic Smyth Ramasamy Uthurusamy Usama M. Fayyad, Gregory Piatetsky Shapiro. Advances in Knowledge discovery and Data Mining, 1996

[8]Manuel Montes y Gómez. Minería de texto: Un nuevo reto computacional. Memoria del 3er. Taller Internacional de Minería de Datos MINDAT-2001, Universidad Panamericana, Ciudad de México, 2001.