Clasificación semántica de textos no estructurados mediante un enfoque evolutivo

Eulalia T. Pacheco-Luz, Felipe Trujillo-Romero y Guillermo Juárez-López

Universidad Tecnológica de la Mixteca,
División de Estudios de Posgrado,
Huajuapan de León, Oaxaca, México

eulalia.pacheco@gmail.com,
{ftrujillo, gjuarezl}@mixteco.utm.mx

Resumen. En la actualidad, cerca del 90% de la información se encuentra plasmada tanto en documentos estructurados como no estructurados. Esto ha dado impulso a la investigación e implementación de diferentes algoritmos para el análisis y clasificación de textos de acuerdo a su orientación semántica. Por ello, en el presente trabajo se describe una manera de clasificación de textos no estructurados mediante el uso de algoritmos evolutivos. Esta técnica será utilizada en el análisis de documentos para determinar la clasificación de acuerdo al enfoque semántico de las palabras que contiene. Para este trabajo se analizaron textos pertenecientes a cuatro géneros literarios diferentes: ciencia-ficción, drama, comedia y terror. Se realizaron varias pruebas obteniendo un desempeño aceptable del sistema implementado.

Palabras clave: minería de textos, clasificación de textos, tesauro, algoritmos genéticos.

1. Introducción

La minería de textos es una disciplina que permite la extracción de información relevante de cantidades extensas de textos. Esto permite definir objetos y sus relaciones, revelando información semántica significativa. El tipo de texto puede ser obtenido de documentos estructurados, es decir que tengan un orden preestablecido en la organización de su contenido, o de no estructurados en los cuales el contenido o información no tiene ningún tipo de orden o estructura.

Adicionalmente la minería de textos se apoya en las técnicas de categorización de texto, procesamiento de lenguaje natural, aprendizaje automático, extracción y recuperación de la información. Existen diferentes opciones tanto comerciales como de software de código abierto. Ejemplo de ello es el software desarrollado por IBM SPSS [3], que es comercial pero ofrece una amplia y sólida variedad de soluciones en minería de textos. Dentro de las herramientas más populares de código abierto, se tiene al lenguaje de *R* [4] y a *RapidMiner* [5]. Este último posee una eficiente interfaz

de usuario, es altamente escalable debido a que maneja clústers y una programación orientada a bases de datos.

Los sistemas de minería de datos permiten el análisis léxico de los textos, especialmente la construcción automática de estructuras de clasificación y categorización que se codifica en forma de tesauros. Algunos ejemplos del uso de este tipo de sistemas se comentan a continuación. En [6] se utilizó RapidMiner (RM) para realizar el análisis de la similitud entre documentos de texto con los contenidos mínimos de los planes de estudio de las Licenciatura en Computación de la Universidad de San Juan. También se utilizó para procesar títulos bibliográficos de la biblioteca, midiendo la similitud sintáctica de los mismos con los contenidos de las diferentes carreras. Por su parte, en [7] se ha aplicado un coeficiente de legibilidad llamado Flesch-Kinkaid, para evaluar el contenido de los discursos del Rey de España y ha llegado a la conclusión de que la complejidad media de los mismos es bastante elevada. Siendo esta similar a la de un artículo científico, con un coeficiente en torno a 50. El estudio se realizó mediante el análisis de frecuencias de aparición de palabras, todo este estudio ha sido realizado con R y el uso de la librería de minería de textos *tm* [9]. En [10] Wei Zong *et al.* proponen un método para la categorización de texto el cual selecciona las características de los documentos basados en la medida de poder discriminativo y de la similitud entre las características usando para ello Maquina de Vectores de Soporte SVM (Support vector machine)

Por su parte, Yuen-Hsien Tseng [11] presenta una metodología de minería de textos especializados para el análisis de patentes mediante un enfoque de distribución de frecuencias de las palabras extraídas de los documentos analizados. En [12], se proponen dos nuevos algoritmos de agrupamiento de texto llamados: 1) Clustering Basado en Secuencias de Palabras Frecuentes (CFWS) y 2) Agrupación en Clústeres Basados en Significado de Secuencias de Palabras Frecuentes (CFWMS). En estos cada documento se reduce a sólo las palabras frecuentes para explorar la secuencia de palabra mediante la construcción de la estructura de un árbol del sufijo generalizado (GST). Finalmente comentamos el trabajo desarrollado por Zelai *et al.* [13] quienes proponen un sistema multiclasificador para categorización de documentos el cual utilizó el algoritmo de clasificación K-NN y un esquema de votación Bayesiano.

Por otro lado, los algoritmos genéticos (AGs) combinan las nociones de supervivencia, del más apto con un intercambio estructurado y aleatorio de características entre individuos de una población de posibles soluciones, conformando un algoritmo de búsqueda aplicado para resolver problemas de optimización en diversos campos [1, 8]. De tal forma, que los algoritmos genéticos se presentan como una herramienta de gran interés para extraer el significado de la información no estructurada de los datos de las organizaciones [2].

Además los algoritmos genéticos presentan ventajas con respecto a otras técnicas entre ellas: 1) no necesitan conocimientos específicos sobre el problema que intentan resolver, 2) operan de forma simultánea con varias soluciones en vez de trabajar de forma secuencial como las técnicas tradicionales, 3) cuando se usan para problemas de optimización-maximizar una función objetivo resultan menos afectados por los máximos locales que las técnicas tradicionales, 4) resulta sumamente fácil ejecutarlos

en las modernas arquitecturas masivas en paralelo y 5) usan operadores probabilísticos en vez de los típicos operadores determinísticos de las otras técnicas.

Entre las aplicaciones que tienen estos algoritmos relacionadas con la clasificación de documentos se puede mencionar a el agrupamiento de documentos y términos, la indización de documentos mediante el aprendizaje de los términos relevantes para describirlos por sus pesos y aprendizaje automático de los pesos de los términos proporcionados previamente por el usuario o de la composición completa de la consulta, incluyendo los términos y los operadores booleanos [14, 15, 16, 17, 18].

Si bien es cierto que los algoritmos genéticos están orientados a la optimización y pueden ser usados para la minería de textos, es necesario realizar las adecuadas modificaciones que permitan su empleo en la categorización de documento. Por ejemplo en el caso de clasificación de textos, cada cromosoma de la población está ligado a un tipo específico de clasificación de artículos como espacio de soluciones.

2. Metodología

Para lograr el descubrimiento de conocimiento, es necesario realizar un proceso que permita tratar la información y finalmente realizar la visualización de los resultados, en la figura 1, se muestra la metodología de la minería de textos que se desarrolló para ser utilizada en el presente trabajo.

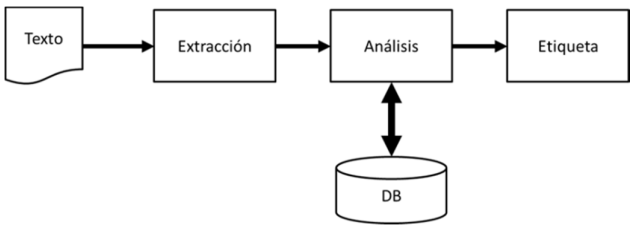


Fig. 1. Metodología de Minería de Textos.

En la misma figura 1, se puede visualizar como primera fase la definición del texto, es decir, se determinó el conjunto de documentos para el posterior análisis y clasificación, así también en esta fase se configuró el tesauro de términos especializados correspondiente a cuatro géneros literarios: ciencia-ficción, drama, comedia y terror. Todo el desarrollo de estudio en cuestión se basó en la norma ISO O 25964-1:2011, esta norma establece las directrices para el establecimiento y el desarrollo de tesauros monolingües [15], y define a tesauro como "*un vocabulario controlado y dinámico, compuesto por términos que tienen entre ellos relaciones semánticas y genéricas y que se aplica a un dominio particular del conocimiento*".

En la fase de extracción o pre-procesamiento se realizó operaciones de transformación sobre el documento, en información estructurada que facilitó su posterior análisis. El análisis de texto consistió en encontrar la secuencia de términos con el objetivo de encontrar patrones de lenguaje, las características que cumplieron dichos

términos se determinaron basándose en la técnica de algoritmos genéticos y la extracción de términos para su categorización.

Una vez terminada la fase de pre-procesamiento, se siguió con la fase de análisis, el cuál consistió en el descubrimiento de conocimiento, para lo cual se aplicó la fase de selección y mutación de algoritmos genéticos, para determinar cuáles son los cromosomas (términos) más representativos y que mayor información semántica proporcionan, a su vez se comparó con la información del tesauro predefinido.

La última fase, fue la visualización de los resultados, en la cual proporciona un ambiente para la exploración de los datos guiados para el usuario que sea lo más amigable posible. Las últimas tendencias presentan los resultados mediante graficas o páginas Web. Una vez obtenidos los conceptos, los términos o las tendencias, se pueden utilizar métodos automáticos de visualización o bien pueden interpretarse los resultados directamente. En este caso los resultados serán las gráficas de agrupación de términos para identificar la clasificación del documento analizado y determinar según la interpretación semántica a que área o campo de aplicación pertenece.

3. Análisis y discusión de resultados

Apoyándose en la metodología de minería de textos y el método de algoritmos genéticos, se desarrolló un programa en el entorno R, el cual permite analizar documentos en formato PDF (Portable Document Format), y clasificarlo en el campo de la literatura al cual pertenece, para la definición de la base de datos del tesauro se utilizó el gestor de base de datos MySql, en las secciones siguientes se describe el proceso realizado.

3.1. Pre-procesamiento

En primer lugar se delimitó el *corpus* a procesar que corresponde a cuatro géneros literarios; drama, ciencia ficción, terror y comedia, a modo de ejemplo se buscaron libros de cada uno de los géneros literarios. Seguidamente se realizó la definición del tesauro; esto incluye los nombres de los campos de especialización así como las palabras técnicas acordes a cada uno de los géneros que se están evaluando.

Para proceder a la extracción de la información, se procesó cada uno de los archivos a fin de convertirlo en archivo de texto plano, lo cual facilitara la extracción de cada uno de los términos.

Una vez que se cuenta con el archivo en texto plano, se procede a la aplicación de la minería de datos, para la limpieza y extracción de cada una de las palabras, para lo cual en primer lugar se instaló la librería *tm* en la herramienta R Studio, esta permitirá delimitar la matriz de términos a explorar, que al final es la matriz de cromosomas. Inmediatamente para realizar la limpieza de la matriz que contiene el texto en crudo, es necesario depurar términos, eliminar los números, los signos de puntuación, palabras auxiliares (artículos, pronombres, etc.), espacios en blanco y se convierte todo en minúsculas. Para realizar este proceso se guarda el texto extraído en un vector y

utilizando las palabras reservadas de la librería *tm* se realiza la limpieza de la información.

Consecuentemente se crea lo que se conoce como una matriz de términos del documento (*m x n*), donde *m* sería el número de descripciones a procesar y *n* sería el número de términos existentes en esos documentos. Los valores de la matriz sería el número de veces que cada fila contiene el término dado. Finalmente se guarda en un *dataframe* las palabras obtenidas para pasar a la etapa de procesamiento.

3.2. Procesamiento

En la etapa anterior, se llevó a cabo la depuración y delimitación de la población de palabras que son el conjunto de individuos a utilizar. Como siguiente paso se tiene que calcular la frecuencia de aparición de las palabras, conjuntamente se establece la conexión con la base de datos, con el objetivo de guardar en una tabla temporal, el conjunto de individuos que componen la población.

El tamaño de la población para este ejemplo es de 20 documentos, los cuales puede aumentar. Así también se definió un clúster de palabras, que es una colección de términos que semánticamente tienen relación entre sí, las cuales serán relacionadas a un campo de la literatura en específico.

El proceso de selección de la población inicial, se genera con los términos que se encuentran en el cuerpo del documento (Ec. 1), los cuales ya fueron guardados en la tabla temporal, cada registro es un cromosoma o individuo y cada uno de ellos está compuesto por un término, el valor del clúster al que pertenece el término, y una probabilidad de aparición asociada (Ec. 2). Es decir, se divide la adaptación de cada uno entre la suma de la de toda la población, y se asocia dicha distribución a una ruleta, dando más espacio en la misma a aquellos individuos que presenten mayor probabilidad de selección, en otras palabras, los mejor adaptados. La longitud el individuo siempre será variable, esto dependerá del texto que se esté analizando.

$$P_i = C_1, C_2, C_3, \dots C_n, \quad (1)$$

$$C_n = v_i, t_x, f, \quad (2)$$

donde:

P_i: Población inicial
$C_1 \dots C_n$: Cromosomas
v_i: Valor del clúster.
t_x: Término
f: Probabilidad.

Para obtener el campo *valor del clúster* (v_i), se realiza un comparación de cada una de las palabras del documento que resultaron del pre-procesamiento, con el tesauro especializado y se anota en el campo el valor del área al cual está asociado.

Pero para cumplir con el objetivo anterior, antes es necesario determinar la entropía de la población (Ec.3), ya que pueden existir diversas palabras con un elevado número de frecuencia, sin embargo semánticamente no aportan información relevante. Por ejemplo las palabras siguientes: que, luego, cuando, donde.

$$H(\mathrm{x}) = -x_i \log_2(x_i) \tag{3}$$

Una vez calculada la entropía se procede a le eliminación de aquellos términos superiores a la ponderación media de la entropía de la población (Ec. 4), así como los términos cuya frecuencia es igual a uno, puesto que tampoco proporcionan información relevante.

$$P = P_i - \left(\frac{H(x)}{n}\right) \qquad \left\{H(x) < \left(\frac{H(x)}{n}\right) \bigwedge P_i > 1\right\} \tag{4}$$

Para ejemplo se analizó el libro titulado "*Adiós Tierra*" del autor Álvaro Cotes Córdoba, del cual seleccionados los términos más representativos, la población para este ejemplo queda definida como se muestra en la Tabla 1.

Logrando de esta manera tener un cromosoma que está representado por los id de las áreas (v_i) a los que pertenece cada término. La representación de este cromosoma o individuo se puede observar en la Figura 2. Este individuo está compuesto por 18 genes.

Tabla 1. Términos representativos del libro "*Adiós Tierra*".

ID AREA (v_i)	TERMINO (t_x)	FRECUENCIA (f)
1	MERCURIO	4
1	TELESCOPIO	4
1	UNIVERSO	4
1	VIDA	4
4	PERSONAJE	4
1	ASTRO	3
1	DIOS	3
1	NASA	3
2	PROHIBIDO	3
1	AÑO	2
1	ATMOSFERA	2
1	DIGITAL	2
1	FISICA	2
1	FUERZA	2
1	PERIODO	2
2	SUDOR	2
3	RISAS	2
3	COLOR	2

1	2	.	.	.									14	15	16	17	18
1	1	1	1	4	1	1	1	2	1	1	1	1	1	1	2	3	3

Fig. 2. Representación del cromosoma a partir de la Tabla 1.

En este trabajo no se aplica el cruce ni la selección de los individuos debido a que todos pasan a la siguiente etapa que es la mutación del cromosoma. Para realizar la

mutación de los genes de los cromosomas, se utilizó el operador de mutación basado en el desplazamiento que describe Michalewizc [16]. Este proceso comienza seleccionando una subcadena de genes de un individuo al azar. Dicha subcadena se extrae del segmento y se inserta en un lugar aleatorio del individuo al cual se le extrae la subcadena. Por ejemplo, a partir del individuo de la figura 2 se toman los genes 14 al 17 que corresponden a la subcadena mostrada en la Figura 3.

1	1	2	3

Fig. 3. Subcadena tomada del individuo de la Fig. 2.

Después, se selecciona aleatoriamente un punto de inserción en el mismo individuo para insertar la subcadena extraída. En este caso fue el punto de inserción fue el gen número 6. Al insertar la subcadena esta reemplaza a los genes que existían anteriormente en el individuo quedando de la manera que se muestra en la Fig. 4.

También se analizó el libro "*El Sonido del Silencio*" del autor Heydee Cabrera, de su análisis se obtuvo la población que se muestra en la Tabla 2.

1	1	1	1	4	**1**	**1**	**2**	**3**	1	1	1	2	1	1	1	1	3

Fig. 4. Individuo mutado.

Tabla 2. Terminos representativos del libro "*El Sonido del Silencio*".

ID AREA	TERMINO	FRECUENCIA
2	SONIDO	56
2	LUNA	52
2	DEMONIOS	26
2	OSCURO	19
1	SILENCIO	19
1	DIOS	18
1	VIDA	14
2	MIEDO	13
1	TIEMPO	12
1	DIA	8
3	CONFUSIÓN	8
2	EXTRAÑO	6
1	GUSTO	4
1	SOL	4
2	CRIATURA	4
2	PELIGROSO	4

ID AREA	TERMINO	FRECUENCIA
1	AGUA	3
2	MUERTE	3
1	OIDO	3
2	SOMBRAS	3
1	SONAR	3
2	HORROR	3
2	PROHIBIDO	3
1	ESTRELLA	2
1	LUZ	2
2	PACTO	2
1	RAYOS	2
1	TIERRA	2
3	DISTURBIOS	2
1	VIDRIO	2
2	TRISTE	2

A partir de los datos obtenidos en la Tabla 2 se genera el cromosoma representativo que se muestra en la Fig. 5. Como se puede apreciar en la Fig. 2 este individuo es más grande (31 genes) que el generado para el ejemplo anterior.

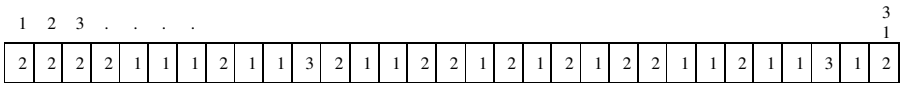


Fig. 5. Representación del cromosoma a partir de la Tabla 2.

A continuación se presenta la mutación del cromosoma mostrado en la Fig. 5. La subcadena extraída para realizar la mutación es [1, 1, 2, 1] y el cromosoma resultante es como se observa en la Fig. 6.

2	2	2	2	1	1	3	2	1	1	2	2	1	**1**	**1**	**2**	**1**	2	1	2	1	2	2	1	1	2	1	1	3	1	2

Fig. 6. Individuo mutado.

Esta mutación continuará hasta contar con un individuo cuyos elementos iguales sean continuos. Esto nos permitirá, posteriormente, dividir el individuo en subgrupos.

3.3. Análisis

Dentro de la información del conocimiento descubierto existen patrones que son más interesantes que otros, razón por la cual, se debe evaluar y determinar cuáles de ellos son los que le aportan conocimiento que le sean de mayor utilidad en la toma de decisiones. Por tal motivo, es necesario realizar la comparación semántica de términos para posteriormente identificar el género al cual pertenece, al contar con un tesauro predefinido es necesario realizar la comparación de términos para identificar cuáles son las palabras que van a definir el área o campo al cual pertenece el documento analizado, puesto que como anteriormente se explicó cada palabra está asociado a un área específica.

Sobre el último individuo cuya longitud siempre será variable para cada población, se determinará el género literario al que pertenece, esto será definido al formar subgrupos con los términos semejantes (v_i), así el subgrupo que mayor cantidad de términos contenga, se compara el número de v_i y se asocia al nombre del género literario que le pertenece, los cuales fueron determinados en el tesauro.

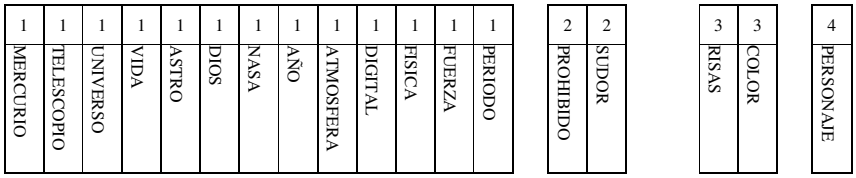


Fig. 7. Agrupación obtenida a partir de la evolución del individuo de la Fig. 2.

A partir de ello se obtuvo el resultado de la agrupación para el libro "*Adiós tierra*" el cual se muestra en la Fig. 7.

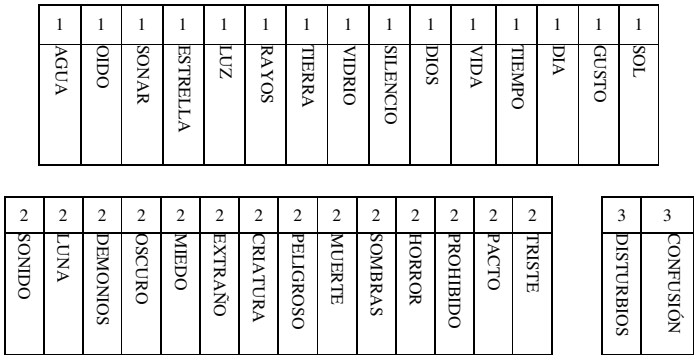


Fig. 8. Agrupación obtenida a partir de la evolución del individuo de la Fig. 5.

En el caso del análisis del libro "*El Sonido del Silencio*" la agrupación queda como se muestra en la Fig. 8.

De la agrupación obtenida se obtiene el nombre del área al cual pertenece el texto analizado mediante la extracción del máximo valor de los índices pertenecientes a los grupos. Esto se realiza mediante la expresión mostrada en la Ec. 5:

$$NA = \max_{i} \sum id\_Area. \tag{5}$$

3.4. Visualización de resultados

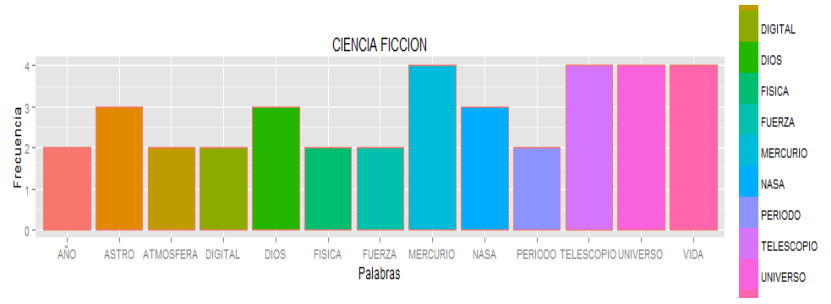
Finalmente, se presenta la visualización de resultados para el usuario final. Las gráficas que se muestran en la Fig. 9 contienen las palabras de mayor referencia semántica contenida en el tesauro, es decir, las palabras que se encontraron en los cromosomas y también se encontraron en el tesauro, así también se visualiza en el encabezado el nombre del campo al que pertenece, dicha información se obtiene seleccionando el id_area cuya mayor frecuencia se presenta en la tabla conceptos, puesto que pueden existir algunas palabras homógrafas.

Esta gráfica cambia de acuerdo al documento analizado puesto que el encabezado, contiene el nombre del área tomado de la base de datos.

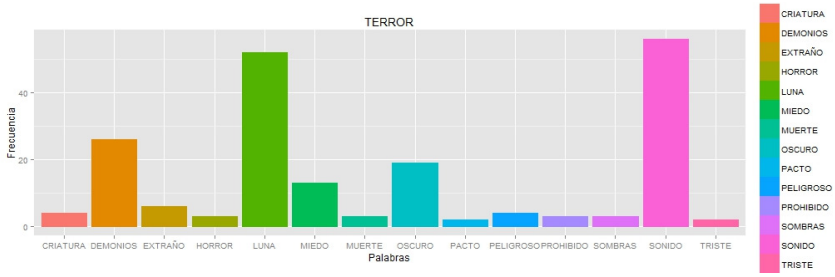
A continuación se muestra las gráficas de los dos ejemplos, donde se puede notar que las palabras que aparecen en cantidad no son elevadas pero sin embargo estos términos están asociados semánticamente y se puede notar la diferencia entre ambos géneros literarios. Además la cantidad de palabras, varía según el documento analizado, sin embargo el título de clasificación, que es el título de la gráfica, siempre se determinara en base al mayor grupo de palabras que se hayan encontrado.

Para la clasificación del documento y lo que nos da la certeza a que grupo pertenece al final es el grupo al que está asociado cada uno de los individuos que ya fue-

ron seleccionados, así por ejemplo, al final de todo el algoritmo tenemos la palabra "Peligroso" y dentro de su estructura de cromosoma está asociado al clúster 2, sabemos que la clasificación es de "Terror."



(a)



(b)

Fig. 9. Gráficas de conceptos.

4. Conclusiones

Actualmente existen varias herramientas para el análisis de documentos no estructurados, ninguno especializado en artículos literarios. Con el desarrollo de una herramienta utilizando el entorno de programación R y las técnicas de minería de datos con la lógica de algoritmos genéticos, se determinó que los algoritmos genéticos no pueden ser aplicados directamente en el proceso de análisis de textos, es necesario realizar algunos cambios al algoritmo para adaptarlo, como lo es el proceso de selección pues además de la probabilidad se toma en cuenta la frecuencia de aparición.

Además se descubrió que la mayor parte de los términos contenidos en los documentos lo podemos considerar basura puesto que no aportan información relevante, y el campo de clasificación es determinado por el mayor conjunto de palabras que se forman en base a su contenido semántico.

Como un trabajo futuro se pretende extender este trabajo para obtener un sistema que sea capaz de clasificar un mayor número de géneros orientándolo hacia el análisis de textos científicos.

Referencias

1. Holland, J. H.: Adaptation in Natural and Artificial Systems. University of Michigan Press, Ann Arbor. Republished by the MIT press (1992)
2. Moore, C.: Diving into data, Info world. http://www.infoworld.com/article/02/10/25/021028feundata_1.html
3. Sitio Web de Soluciones y software de analítica predictiva (Software SPSS). IBM, http://www-01.ibm.com/software/mx/analytics/spss
4. Venables, W. N., Smith, D. M.: the R Core Team. Introduction to R, version 3.1.3, R Core Team (2015)
5. Sitio web de RapidMiner. RapidMiner Studio. https://rapidminer.com
6. Gutiérrez Mag, L.: Pertinencias de planes de estudio de carreras de informática con normativas establecidas por CONEAU. Universidad Nacional de San Juan (2013)
7. Serrano Sánchez, A.: Minería de textos o cómo analizar los discursos del Rey. Universidad Francisco de Victoria. http://ti3.ceiec.es/mineria-de-textos-o-como-analizar-los-discursos-del-rey
8. Goldberg, D.: Genetics Algorithms in Search, Optimization and Machine Learning. Addison Wesley (1989)
9. Feinerer, I., Hornik, K., Meyer, D.: Text Mining Infrastructure in R, Journal of Statistical Software, vol. 25 (5), pp. 1548–7660 (2008)
10. Zong, W., Wu, F., Chu, L.K., Sculli, D.: Discriminative and Semantic Feature Selection Method for Text Categorization. International Journal of Production Economics, available online, ISSN 0925-5273 (2015)
11. Tseng, Y., Lin, C., Lin, Y.: Text Mining Techniques for Patent Analysis. Information Processing and Management, vol. 43 (5), pp. 1216–1247 (2007)
12. Li, Y., Chung, S. M., Holt, J. D.: Text Document Clustering based on Frequent Word Meaning Sequences. Data & Knowledge Engineering, vol. 64 (1), pp. 381–404 (2008)
13. Zelai, A., Alegria, I., Arregi, O., Sierra, B.: A Multiclass/Multilabel Document Categorization System: Combining Multiple Classifiers in a Reduced Dimensión. University of the Basque Country, UPV-EHU, Computer Science Faculty, Euskal-Herria, Spain (2011)
14. Mukherjee, I., Al-Fayoumi, M., Mahanti, P.K., Jha, R., Al-Bidewi, I.: Content Analysis based on Text Mining using Genetic Algorithm. 2nd International Conference on Computer Technology and Development (ICCTD 2010), pp. 432–436 (2010)
15. Khalessizadeh, S. M., Zaefarian, R., Nasseri, S.H., Ardil, E.: Genetic Mining: Using Genetic Algorithm for Topic based on Concept Distribution. Proceedings of World Academy of Science, Engineering and Technology, vol. 13, ISSN 1307-6884 (2006)
16. Yolis, E., Britos, P., Sicre, J., Servetto, A., García-Martínez, R., Perichinsky. G.: Algoritmos genéticos aplicados a la categorización automática de documentos. IX Congreso Argentino de Ciencias de la Computación (CACIC), La Plata, Argentina (2003)
17. Bharadwaj, D., Shukla, S.: Text Mining Technique using Genetic Algorithm. Proceedings on International Conference on Advances in Computer Application (ICACA) (2013)
18. Shivani, P., Gandhi, P.: A Detailed Study on Text Mining using Genetic Algorithm, International Journal of Engineering Development and Research (IJEDR), ISSN:2321-9939, vol. 1 (2), pp. 108–113 (2014)
19. Sitio Web de la Organización internacional para la estandarización. http://www.iso.org/iso/home/store/catalogue_ics/catalogue_detail_ics.htm?csnumber=53657
20. Michalewicz, Z.: Genetic Algorithms + DataStructures = Evolution Programs. Springer-Verlag, BerlinHeidelberg (1992)