Desarrollo de un sistema para medir similitud entre clases

R. Guzmán-Cabrera, J.C. Ruiz, M. Torres-Cisneros

University of Guanajuato, Engineering Division, Irapuato-Salamanca, Guanajuato, México

guzmanc@ugto.mx

Resumen. El incremento continuo de información en formato digital obliga a contar con nuevos métodos y técnicas para acceder, recopilar y organizar estos volúmenes de información textual. Una de las técnicas más utilizadas para organizar dicha información es la clasificación de documentos. Los sistemas de clasificación automática de textos tienen una baja eficiencia cuando las clases son muy parecidas, y en este caso es muy importante el poder identificar aquellos atributos que nos permiten separar una clase de otra. En este trabajo se presenta un sistema para generar gráficas de similitud entre documentos pertenecientes a clases de un corpus dado, tarea previa al proceso de clasificación automática. Estas gráficas son utilizadas como un método de refinamiento auxiliándose de las similitudes entre los documentos no clasificados. Con esto se busca poder anticipar el desempeño de un método de clasificación automática. Los resultados obtenidos permiten ver la viabilidad de la metodología propuesta.

Palabras clave: Clasificación de documentos, similitud.

1. Introducción

En la actualidad el almacenamiento de contenido digital se ha vuelto más abundante y menos costoso. Esto ha provocado que la cantidad de información digital generada por compañías de diferentes rubros crezca a una gran velocidad, generando de esta forma, grandes repositorios de conocimiento. Sin embargo, esto ha provocado la necesidad de crear técnicas para poder clasificar de manera automática estos volúmenes de datos [1].

En el caso de clasificación documentos de texto, los documentos son convertidos de su contenido original, a arreglos de información los cuales representan el contenido de esos documentos. Una de las técnicas más utilizadas para la representación de los documentos es la de usar la característica de frecuencia de aparición de una palabra o frase en el documento [1]. Además de la necesidad de ejemplos para el entrenamiento [2], un problema recurrente al que se enfrenta un clasificador, es el de la similitud entre los documentos de diferentes clases, este problema consiste en qué tan parecido es un documento de una clase a con respecto de una clase b. La similitud, por lo general está representada por una escala numérica entre 0 y 1, donde 0 representa que no existe similitud alguna y 1 representa que el documento es el mismo. Sean di y dj documentos

representados de la forma $d_j = (w_{1j}, w_{2j}, ..., w_{mj})$, entonces algunas medidas de similitud o distancia se definen a continuación:

El coeficiente de similitud de Jaccard mide la similitud entre dos conjuntos de muestras. Aunque originalmente fue utilizado para comparar tipos de flores en un ecosistema [3], ha tenido buena aceptación en el campo del análisis de documentos [4].

$$\boldsymbol{jaccard}(\boldsymbol{d_i}, \boldsymbol{d_j}) = \frac{\sum(\boldsymbol{w_{ik}} \times \boldsymbol{w_{jk}})}{(\sum \boldsymbol{w_{ik}^2} + \sum \boldsymbol{w_{jk}^2}) - (\sum \boldsymbol{w_{ik}} + \sum \boldsymbol{w_{jk}})}$$

El coeficiente de Dice determina la similitud entre dos documentos pesa- dos dando importancia a los atributos de la intersección [5].

$$\boldsymbol{Dice}(\boldsymbol{d_i}, \boldsymbol{d_j}) = \frac{\boldsymbol{2}\sum(\boldsymbol{w_{ki}} \times \boldsymbol{w_{kj}})}{\sum \boldsymbol{w_{ki}} + \sum \boldsymbol{w_{kj}}}$$

La medida coseno es una de las más populares para determinar la similitud de los documentos. El objetivo es determinar el ángulo entre dos vectores, en este caso los vectores de representación de los documentos [6].

$$\boldsymbol{Coseno}(\boldsymbol{d_i}, \boldsymbol{d_j}) = \frac{\sum \boldsymbol{w_{ki}} \times \boldsymbol{w_{kj}}}{\sqrt{\sum \boldsymbol{w_{ki}^2}} \times \sqrt{\sum \boldsymbol{w_{kj}^2}}}$$

2. Desarrollo

En la figura 1 se muestra el proceso que realizará el programa en un esquema general.

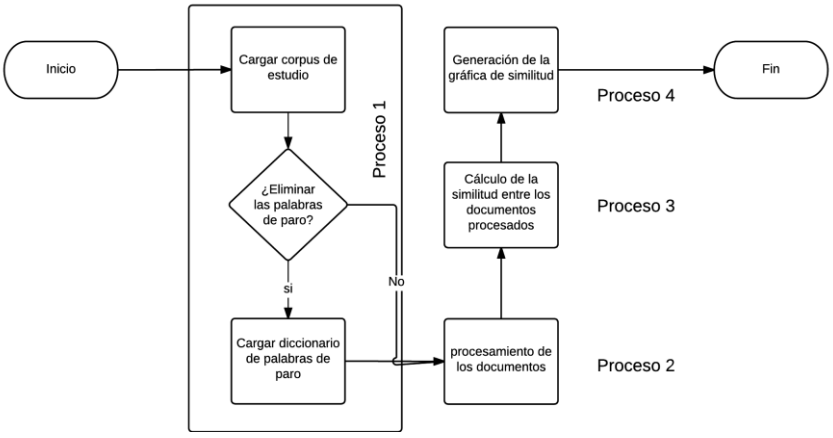


Fig. 1. Diagrama de flujo del proceso del programa.

En la figura 2, se muestra la ventana principal de la aplicación.

Una función adicional de la ventana principal, es el desplegar el histograma de frecuencias referente a cada clase o a cada documento, este módulo se muestra ante el evento de dar doble clic en alguna de las clases o alguno de los documentos. En la figura 3 se muestra un histograma de palabras a manera de ejemplo. El módulo de

visualización de histogramas de frecuencia, tiene como finalidad entregar una asistencia visual al usuario para el análisis de las tablas de frecuencias entregadas por el programa. Cuenta con la operación de acercar/alejar y funciona tanto para mostrar los histogramas de las clases, como de los documentos únicos.

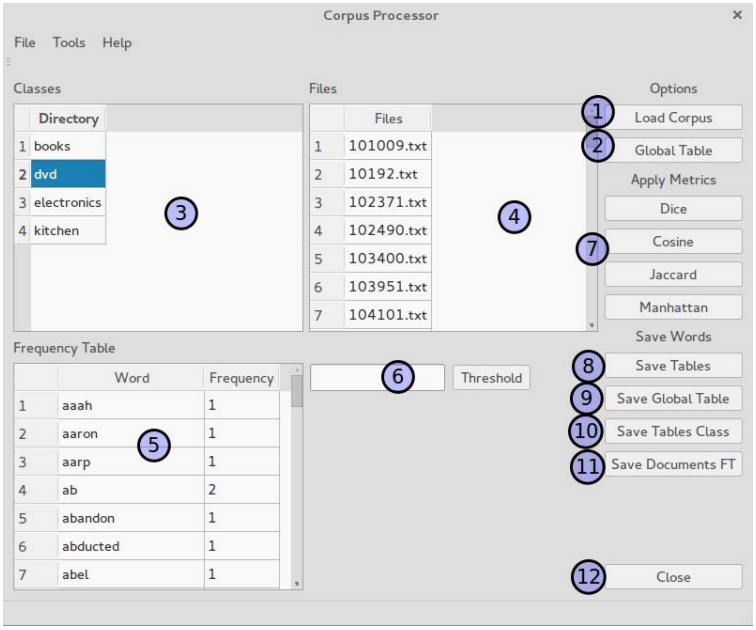


Fig. 2. Ventana principal de la aplicación.

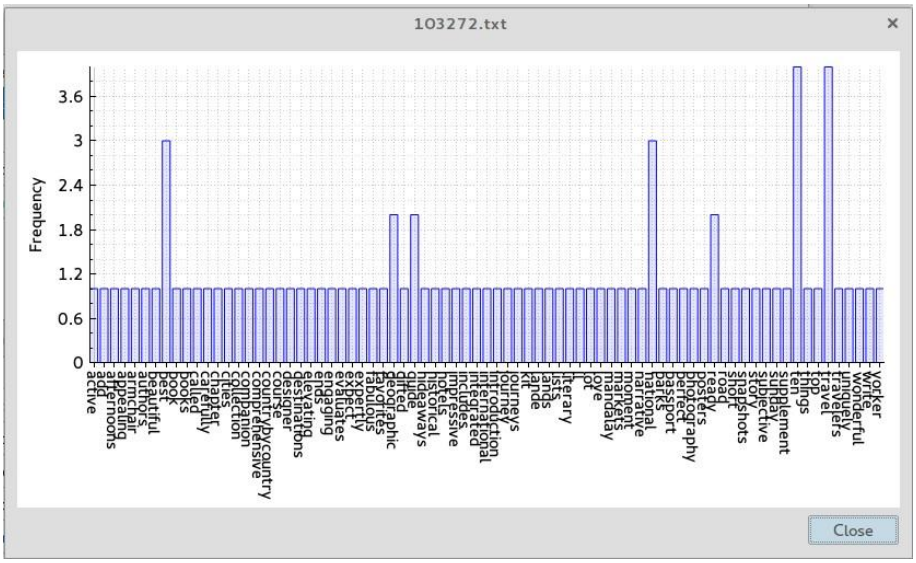


Fig. 3. Módulo de visualización de histogramas.

3. Pruebas y resultados

El corpus de estudio utilizado para las pruebas es el "Multi-Domain Sentiment Dataset (version 2.0)" [7]. El corpus está compuesto de opiniones de 4 diferentes categorías de productos tomados de la base de datos de Amazon [8]. Las cuatro categorías seleccionadas son:

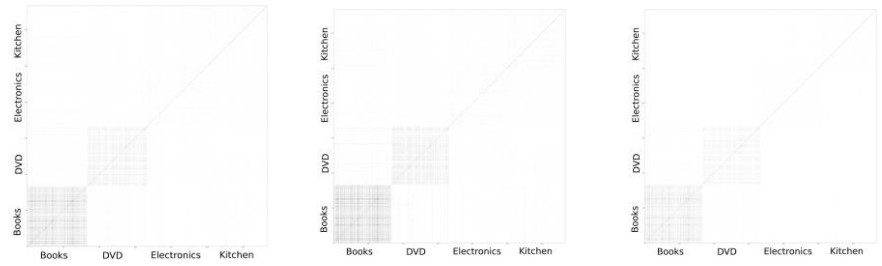
- Libros: 1463 opiniones positivas y 1039 opiniones negativas.
- DVD'S: 1391 opiniones positivas y 1396 opiniones negativas.
- Artículos de cocina: 832 opiniones positivas y 922 opiniones negativas.
- Electrónicos: 948 opiniones positivas y 1014 opiniones negativas.

Para la realización de las pruebas se creó un conjunto de documentos compuesto de 1600 archivos (400 archivos en cada clase). Este conjuntos a su vez, se subdividió en un conjunto de entrenamiento (70 % del conjunto original) y un conjunto de prueba (30 % del conjunto original) los cuales se utilizaron para realizar el proceso clasificación automática de documentos. Los resultados de similitud obtenidos para este conjunto se muestran en la figura 4.

	books	dvd	electronics	kitchen
kitchen	0.0681353	0.142873	0.420928	1
electronics	0.0500725	0.153189	1	0.420928
dvd	0.110637	1	0.153189	0.142873
books	1	0.110637	0.0500725	0.068135

	books	dvd	electronics	kitchen
kitchen	0.139786	0.176653	0.430826	1
electronics	0.124711	0.218162	1	0.430826
dvd	0.13653	1	0.218162	0.176653
books	1	0.13653	0.124711	0.139786

	books	dvd	electronics	kitchen
kitchen	0.0352692	0.0769325	0.266567	1
electronics	0.0256792	0.0829481	1	0.266567
dvd	0.0585581	1	0.0829481	0.076932
books	1	0.0585581	0.0256792	0.035269



(a) Dice (b) Coseno (c) Jaccard

Fig. 4. Similitud entre clases resultante de cada métrica utilizando un umbral de frecuencia de 2.

Entre mayor es el valor del umbral de la frecuencia, mayor es la eliminación de palabras en los documentos, llegando a un punto en el cual comienza a existir pérdida de información; por lo que se espera que el clasificador actúe mejor cuando cuenta con una cantidad de información relevante sin caer en la pérdida de ésta.

Para las pruebas de clasificación se utilizó el software WEKA de la Universidad de Waikato [9] el cual provee de diferentes técnicas de clasificación. Los métodos de clasificación seleccionados para las pruebas fueron, Naive Bayes y Máquinas de Vectores de Soporte.

En la tabla 1 se presentan los resultados obtenidos por los clasificadores bajo diferentes umbrales de frecuencia para la prueba con 1600 documentos. Como se puede apreciar, la precisión del clasificador disminuye, conforme aumenta la cantidad de palabras descartadas por el umbral de frecuencia (TH).

Tabla 1. Resultados clasificación de 1,600 documentos.

Clasificador	Entrenamiento	Prueba	Precisión	Recuerdo	F-Measure
Naive Bayes Original	280	120	0.834	0.827	0.826
SVM Original	280	120	0.822	0.802	0.806
Naive Bayes TH2	280	120	0.725	0.684	0.691
SVM TH2	280	120	0.735	0.697	0.705
Naive Bayes TH3	280	120	0.739	0.602	0.619
SVM TH3	280	120	0.747	0.623	0.642
Naive Bayes TH4	280	120	0.734	0.604	0.603
SVM TH4	280	120	0.76	0.619	0.618

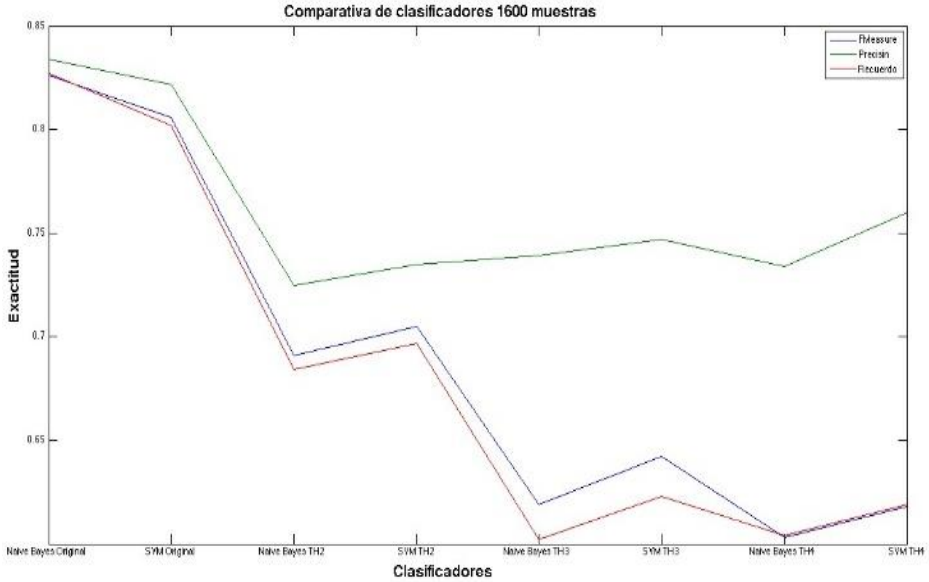


Fig. 5. Exactitud de los clasificadores para 1,600 documentos.

5. Conclusiones y trabajo futuro

Con el desarrollo de este proyecto se demostró que el resultado presentado por las gráficas de similitud, nos permite saber de una manera previa, cómo se comportará el clasificador. A su vez se pudo apreciar cómo el filtrado de información mediante un

umbral de frecuencia puede contribuir a un mejor desempeño, siempre y cuando no se llegue a un grado donde exista una pérdida de información relevante.

Para las métricas de Cosenos, Dice y Jaccard se puede apreciar que la métrica con mejores resultados fue la de Cosenos al mostrar una mejor nitidez ante las otras bajo el mismo conjunto de información.

Referencias

1. Chen, Z., Ni, C., Murphey, Y.L.: Neural network approaches for text document categorization. In: International Joint Conference on Neural Networks (IJCNN'06), pp. 1054–1060 (2006)
2. Nigam, K., McCallum, A.K., Thrun, S., Mitchell, T.: Text classification from labeled and unlabeled documents using EM. Mach. Learn. 39(2-3), pp. 103–134 (2000)
3. Jaccard, P.: Étude comparative de la distribution florale dans une portion des Alpes et des Jura. Bulletin del la Société Vaudoise des Sciences Naturelles 37, pp. 547–579 (1901)
4. Niwattanakul, S., Singthongchai, J., Wanapu, S.: Using of Jaccard coefficient for keywords similarity. In: Proceedings of the International MultiConference of Engineers and Computer Scientists, Vol. 1, IMECS, March 13-15, Hong Kong (2013)
5. Salton, G., McGill, M.: Introduction to Modern Information Retrieval. McGraw Hill (1983)
6. Frakes, W.B., Baeza-Yates, R.: Information Retrieval: Data structures and Algorithms. Englewod Cliffs, Prentice Hall (1992)
7. Multi-domain sentiment dataset (version 2.0). http://www.cs.jhu.edu/mdredze/datasets/ sentiment/
8. Amazon site. http://www.amazon.com
9. Weka 3: Data mining software in java. http://www.cs.waikato.ac.nz/ml/weka/