

Evaluación del *clustering* de páginas web mediante funciones de peso y combinación heurística de criterios

Arantza Casillas Rubio

Dpto. Electricidad y Electrónica
(UPV-EHU)
Apto. de correos 664. 48940
Bilbao
arantza@we.lc.ehu.es

Víctor Fresno Fernández

Dpto. de Informática, Estadística
y Telemática (URJC)
c/Tulipán s/n 28933
Móstoles - Madrid
victor.fresno@urjc.es

Raquel Martínez Unanue

Dpto. de Informática, Estadística
y Telemática (URJC)
c/Tulipán s/n 28933
Móstoles - Madrid
raquel.martinez@urjc.es

Soto Montalvo Herranz

Dpto. de Informática, Estadística
y Telemática (URJC)
c/Tulipán s/n 28933
Móstoles - Madrid
soto.montalvo@urjc.es

Resumen: El *clustering* de páginas web facilita, entre otras tareas, la valoración y búsqueda de resultados de un buscador de páginas web. Uno de los aspectos clave del proceso de *clustering* es la función de peso que se aplica a los rasgos seleccionados para representar dichas páginas. Este artículo presenta la evaluación de los resultados de un algoritmo de *clustering* de partición sobre una colección de referencia de páginas web, utilizando siete funciones de peso distintas y dos tipos de reducción de rasgos. Se han comparado cinco funciones bien conocidas, basadas únicamente en el contenido textual de las páginas web, con otras dos funciones de peso basadas en una combinación heurística de criterios, entre los que destaca la utilización de la información de las anotaciones HTML. Estas dos últimas han sido propuestas, por parte de uno de los autores, en trabajos anteriores. Se ha comprobado que los mejores resultados se obtienen con la función que combina en forma borrosa este tipo de criterios.

Palabras clave: *clustering* de páginas web, funciones de peso, representación de páginas web, combinación borrosa de criterios

Abstract: Web page clustering can help in the evaluation and search of the results of search engines, among other things. The different term weighting functions applied to the selected features to represent web pages is a main aspect in clustering task. In this paper, seven different term weighting functions are evaluated by means of the results of a partitioning clustering algorithm, with a reference web page collection. In addition, two feature reduction methods are applied. Five of them are well-known term weighting functions from text content analysis; the other two are based on a heuristic criteria combination, which consider HTML mark-up information. These two representations have been proposed in previous works by one of the authors. We have verified that the best results are obtained when the term weighting function based on a fuzzy criteria combination is used.

Keywords: web page clustering, term weighting functions, web page representation, Fuzzy Combination of criteria

1. Introducción

Internet se ha convertido en una enorme fuente de información y su utilidad dependerá de la habilidad de las herramientas de extracción y recuperación de información, así como de la posibilidad de descubrir relaciones, anteriormente desconocidas, entre los

datos de la web ((Zhongmei y Ben, 2003), (Jiayun, 2004), (Iavernaro, 2004)).

El acceso a la información web se realiza principalmente mediante motores de búsqueda y directorios web temáticos. En (Gonzalo, 2004) se plantea que usar un motor de búsqueda tipo *google* funciona bien cuando buscamos información como páginas per-

sonales o sitios correspondientes a instituciones, corporaciones o eventos. Sin embargo, si queremos encontrar un formulario particular de una institución, explorar de manera transversal, este modelo presenta algunas limitaciones. La lista ordenada que se nos devuelve no está conceptualmente organizada, ni relaciona información extraída de diferentes páginas. Sin embargo, hay otros buscadores, como *Vivísimo*, que además de la lista de documentos relevantes, presenta la jerarquía de *clusters* asociada. En el caso de los directorios web, los documentos se presentan clasificados en taxonomías y la búsqueda se fundamenta en dicha estructuración.

En este contexto, técnicas de *clustering* de documentos son muy útiles ya que se pueden aplicar a tareas como: agrupación automática previa y posterior a la búsqueda, búsquedas por similitud, visualización de los resultados de una búsqueda de manera estructurada ((Good, 1958), (Fairthorne, 1961) y (Needham, 1961)).

En cualquier proceso de *clustering* hay dos aspectos clave: el propio algoritmo y la función de peso aplicada en la selección de rasgos de las páginas. En este trabajo se evalúa, para un algoritmo conocido concreto, la adaptación de siete funciones de peso diferentes, aplicando dos métodos de reducción del número de rasgos. Se trata de comprobar si, usando información extraída del etiquetado HTML como parte de la representación de una página web, se mejoran los resultados del *clustering* frente a funciones clásicas en el ámbito del *clustering* de textos planos. Cinco de las funciones para calcular el peso se basan únicamente en el contenido textual de la página web, y otras dos funciones tienen en cuenta además del texto, la información de las etiquetas HTML (énfasis y las etiquetas de título *title*) y la posición de los rasgos. La experimentación se realiza sobre una colección de referencia para evaluar *clustering* de páginas web.

El resto del artículo se organiza como sigue: en la sección 2 se resumen las principales aproximaciones existentes en la representación de páginas web. La sección 3 muestra las diferentes funciones de peso que hemos estudiado. En la sección 4 se describe la colección usada para la evaluación. La sección 5 presenta los dos tipos de reducción de rasgos, los experimentos, el algoritmo de *clustering*, el método de evaluación y los resultados. Por

último, la sección 6 incluye las conclusiones.

2. Representación de páginas web

La representación de páginas web varía en función de los elementos de la página que se tengan en cuenta: texto plano, texto enriquecido con etiquetas HTML, metacontenidos, enlaces, texto asociado a enlaces ...

Inicialmente, para la representación de páginas web se aplicaron directamente las mismas técnicas que se aplicaban a los textos. Estas representaciones se incluirían dentro de la *representación por contenido*, que no tienen en cuenta ni la estructura ni la topología del documento.

Posteriormente, se introdujeron elementos propios de la web en la representación, como los hiperenlaces e información asociada a las estructuras de grafo que forman ((Getoor, 2003), (Glover et al., 2002), (Chakrabarti, 2002), (Asirvatham y Ravi, 2001) y (Lu y Getoor, 2003)) combinando información interna de la página con información presente en otras páginas que apuntan a ella (*authorities*) o apuntadas por ella (*hubs*), y las propias URLs (Merkl, 1998). Estas son las ideas que subyacen en los algoritmos PageRank (Brin y Page, 1998) que usa *google* y (Barfouroush et al., 2002). Estas representaciones, denominadas *por contexto*, también analizan componentes multimedia ((Guglielmo y Rowe, 1996), (Harmadas, Sanderson, y Dunlop, 1997) y (Srihari, 1995)) o meta-etiquetas. Sin embargo, en (Pierre, 2001) se muestra que el porcentaje de páginas web que contienen meta-etiquetas no es superior al 30%.

De este modo, cabría esperar que el coste relativo de introducir información de contexto frente a considerar únicamente información contenida en la página, resultará cada vez más alto según siga creciendo la Web. Así, el hecho de mejorar las representaciones que no requieran información externa a la propia página podría tener gran interés.

3. Funciones de peso

Las páginas web se han representado usando el modelo de espacio vectorial (Salton y McGill, 1983). En este modelo, cada página web se representa mediante un vector, donde cada componente representa el peso de un rasgo de la colección en la página. Por

tanto, usando diferentes funciones de peso, se obtendrán diferentes representaciones.

En primer lugar, usamos cinco funciones de peso bien conocidas, basadas únicamente en el texto plano de las páginas web.

Binaria (Binary, B): el peso es un valor binario $\{0,1\}$ en función de si el rasgo aparece o no en el documento.

Frecuencia del Término (Term Frequency, TF): cada rasgo tiene una importancia proporcional al número de veces que aparece en el documento (Luhn, 1957).

Frecuencia Inversa del Documento Binaria (Binary Inverse Document Frequency, BinIDF): la importancia de un rasgo, si aparece en un documento, es inversamente proporcional al número de documentos que lo contienen; el factor IDF de un término t es: $IDF(t) = \log \frac{N}{df(t)}$ y el peso del término en BinIDF es: $W(d, t) = B(d, t) \times IDF(t)$, donde $B(d, t) = \{0, 1\}$.

TF-IDF: La combinación de pesos de un término t en un documento d viene dada por: $W(d, t) = TF(d, t) \times IDF(t)$ (Salton y Yang, 1973).

WIDF: es una extensión de IDF que incorpora la frecuencia del término sobre la colección de documentos: $WIDF(d, t) = TF(d, t) \sum_{i \in D} TF(i, t)$ (Salton, 1988).

Además, usamos dos funciones heurísticas que combinan varios criterios, entre los que destaca utilizar información de las anotaciones HTML de enfatizado y título, que pueden reflejar la importancia que da el autor a determinadas partes del contenido. Estas dos representaciones son: la combinación analítica de criterios (*Analytic Combination of Criteria*, ACC) (Fresno y Ribeiro, 2004) y la combinación borrosa de criterios (*Fuzzy Combination of Criteria*, FCC) (Ribeiro et al., 2002), que se describen a continuación.

3.1. Combinación Analítica de Criterios (ACC)

La función de peso ACC combina linealmente criterios según las siguientes funciones:

Función de frecuencia de una palabra en la página web:

$$f_f(i) = \frac{n_f}{N_{tot}} \quad (1)$$

donde $n_f(i)$ es el número de ocurrencias de la palabra i en la página y N_{tot} es el número total de palabras de la página web. Se normaliza la función usando $\sum_1^k f_f(i) = 1$, donde k es el número de palabras diferentes en el documento.

Función de frecuencia de una palabra en el título:

$$f_t(i) = \frac{n_t}{N_{tit}} \quad (2)$$

donde $n_t(i)$ y N_{tit} son el número de ocurrencias de una palabra i en el título y el número total de palabras en el título, respectivamente. Se normaliza mediante $\sum_1^k f_t(i) = 1$, siendo k el número de palabras diferentes en el título.

Función de enfatizado de una palabra:

$$f_e(i) = \frac{n_e}{N_{emph}} \quad (3)$$

donde $n_e(i)$ y N_{emph} son el número de veces que una palabra se ha enfatizado en el documento y el número total de palabras enfatizadas, normalizando $\sum_1^k f_e(i) = 1$.

Función de posición:

Para cuantificar el criterio de la posición, la página web se divide en 4 partes iguales. Siendo $n_{tot}(i) = n_{1,4}(i) + n_{2,3}(i)$, y $n_{1,4}(i)$ y $n_{2,3}(i)$ el número de veces que el término i aparece en la partes primera y última, y segunda y tercera respectivamente. La función es:

$$f_p(i) = \frac{2n_{1,4}(i) + n_t(i)}{2n_{tot}(i) + N_{tot}(i)} \quad (4)$$

Esta expresión es una particularización de una expresión más general (Fresno y Ribeiro, 2004). Finalmente, la combinación de estos criterios para cada rasgo i en la función de peso ACC, viene dada por la siguiente función de relevancia, particularizada en la sección 5:

$$r_i = C_1 f_f(i) + C_2 f_t(i) + C_3 f_e(i) + C_4 f_p(i)$$

3.2. Combinación Borrosa de Criterios (FCC)

La función de peso FCC combina los criterios usando un sistema de reglas borroso para la asignación y combinación de pesos.

Las variables lingüísticas del sistema borroso son: *text-frequency* etiquetadas como “Low”, “Medium” y “High”; *title-frequency* con etiquetas “Low” y “High”; *emphasis* con “Low”, “Medium” y “High”; y *global-position*.

Con el fin de hacer las reglas borrosas independientes del tamaño del documento, las entradas de *text-frequency*, *title-frequency* y *emphasis*, se normalizan con la frecuencia mayor en cada criterio. Por ello, las ecuaciones (1), (2) y (3), que servirán como función de entrada a las variables lingüísticas, se modificarían haciendo el denominador el número máximo, en lugar del total.

Con respecto al criterio “posición”, la variable *global-position* se calcula a partir de un sistema borroso auxiliar. Este sistema tiene como entrada la variable *line-position* y como salida la variable *global-position*, etiquetadas como “Standard” y “Preferential”. El valor de la salida se calcula desde las diferentes posiciones en las que aparece el rasgo.

Como salida del módulo borroso sólo hay una variable por cada rasgo de la página: *relevance*, etiquetada como “NonRelevant”, “LowRelevant”, “MediumRelevant”, “HighRelevant” y “VeryRelevant”. El motor de inferencia está basado en el algoritmo Centro de Masas (COM), que pesa la salida de cada regla contra el grado de verdad de cada antecedente. El conjunto de reglas y la explicación de los detalles del módulo borroso se puede encontrar en (Ribeiro et al., 2002).

El establecimiento de las reglas se basa en las siguientes consideraciones:

1. Una página web puede tener palabras no enfatizadas.
2. Una palabra que aparezca en el título no siempre es relevante (el título puede haber sido generado de forma automática por un editor HTML).
3. En general, con páginas más largas tiene más peso el criterio de la posición que en páginas más cortas.
4. Una palabra con una frecuencia alta en una página, podría significar que ésta es

de propósito general y, por consiguiente, no ayudaría a discriminar.

4. Colección de páginas web: BankSearch

Con el fin de evaluar el resultado del *clustering* con las diferentes funciones de peso y los dos tipos de reducciones, se ha utilizado la colección *BankSearch* (Sinka y Corne, 2002). Esta colección está formada por 11.000 páginas web en inglés preclasificadas en 11 categorías del mismo tamaño. Fue recopilada y clasificada con el propósito de ser utilizada como una colección de referencia para evaluar *clustering* de páginas web.

Para los experimentos se han considerado 10 categorías principales. Las tres primeras pertenecen al tema general “Bancos & Finanzas”: Bancos Comerciales (A), Sociedades de Crédito Hipotecario (B) y Aseguradoras (C). Las tres siguientes se refieren al tema “Lenguajes de Programación”: Java (D), C/C++ (E) y Visual Basic (D). Las dos siguientes corresponden a “Ciencia”: Astronomía (G) y Biología (H). Por último, las dos restantes corresponden al tema “Deportes”: Fútbol (I) y Deportes de Motor (J).

Para representar cada página sólo se han considerado rasgos de longitud no superior a 30 caracteres. El preproceso constó de las siguientes fases:

1. Lematizar el contenido.
2. Eliminar las palabras de una lista de palabras vacías de contenido (artículos, determinantes, ...) que normalmente se utiliza en Recuperación de Información.
3. Eliminar aquellos rasgos que sólo aparecían una vez en la colección.
4. Contar el número de veces que cada rasgo aparece en cada página web y el número de páginas en las que aparece cada rasgo.
5. Registrar la posición de cada rasgo en cada página web y si aparece o no en elementos HTML de enfatizado. Esta información es necesaria para las funciones ACC y FCC.

5. Experimentos

Se ha realizado el *clustering* de la colección descrita con las siete funciones de peso y con dos tipos de reducción de rasgos. Los coeficientes para la combinación lineal de la representación ACC se han estimado en base a una investigación previa (Fresno y Ribeiro, 2004) sobre la influencia de cada criterio en la representación. Esta investigación se

llevó a cabo con una colección de páginas web diferente de la colección empleada en este trabajo. Como resultado, se estimaron los siguientes coeficientes en la combinación lineal de criterios: C_1 (Frequency) 0.30, C_2 (Title) 0.15, C_3 (Emphasis) 0.25, C_4 (Position) 0.30.

5.1. Reducción de Rasgos

La dimensión, es decir, el número de rasgos de la representación de una colección de páginas web es un aspecto crítico. En estos experimentos se han realizado dos tipos de reducción de rasgos:

1. Reducción “Min-Max”: sólo los rasgos que aparezcan más de $FFmin$ veces en más de $DFmin$ páginas web, y menos de $FFmax$ veces en menos de $DFmax$ páginas son seleccionados. Se trata de una variación sobre la reducción clásica *Document Frequency* (Sebastiani, 2002).
2. Reducción “PF”: las propias funciones de peso se han usado como método de reducción. Por lo tanto, se seleccionan los N rasgos más relevantes (aquellos con mayor valor de función) de cada página web. En este caso, se ha podido aplicar esta reducción para todas las funciones de peso, excepto para la función Binaria (B).

Con el fin de obtener diferentes magnitudes de reducción, y probar la calidad de las representaciones para tareas de *clustering*, se han probado diferentes valores para las variables $FFmin$, $DFmin$, $FFmax$, $DFmax$, y N . Así, el número máximo de rasgos es de un orden de magnitud menor que el número de rasgos inicial.

Es importante destacar que las representaciones B, TF, ACC y FCC son independientes de la información de la colección; es decir, sólo necesitan información de la propia página web para representarla. Sin embargo, las representaciones BinIDF, TF-IDF y WIDF necesitan la información de la colección para representar cada página.

5.2. Algoritmo de *Clustering*

El *clustering* de un conjunto de documentos consiste en dividirlo en conjuntos disjuntos de *clusters* (subconjuntos), tales que los documentos pertenecientes al mismo *cluster*

sean “similares” entre sí y sean menos “similares” de los pertenecientes a los demás *clusters*.

Se pretende evaluar qué funciones de peso pueden mejorar el *clustering* de páginas web. Se ha utilizado un algoritmo de *clustering* de partición, en concreto, el algoritmo *Direct* de la conocida librería CLUTO (Karypis, 2002).

5.3. Medidas de Evaluación

Con el fin de evaluar la calidad del *clustering* se ha realizado una evaluación externa, por medio de la medida-F (Rijsbergen, 1974). Denominamos “clases” a los grupos proporcionados por la colección *BankSearch*, y “clusters” a los grupos obtenidos por el algoritmo de *clustering*. Esta medida combina las medidas de precisión y cobertura.

La medida-F del *cluster* j y la clase i viene dada por

$$F(i, j) = \frac{2 \times Recall(i, j) \times Precision(i, j)}{Precision(i, j) + Recall(i, j)} \quad (5)$$

donde $Recall(i, j) = \frac{n_{ij}}{n_j}$, $Precision(i, j) = \frac{n_{ij}}{n_i}$, n_{ij} es el número de miembros de la clase i en el *cluster* j , n_j es el número de miembros del *cluster* j y n_i el número de miembros de la clase i . Para todos los *clusters*: $F = \sum_i \frac{n_i}{n} \max\{F(i, j)\}$, donde n es el número de páginas web. Un mayor valor de la medida-F indica mejor calidad del *clustering*.

5.4. Resultados

Los resultados de la experimentación realizada se pueden ver en las figuras 1 y 2. La primera corresponde a los resultados del *clustering* utilizando como método de reducción la propia función, y la figura 2 se refiere a los resultados mediante la reducción que hemos denominado Min-Max.

Cuando se utiliza la propia función de peso para reducir el número de rasgos, la función FCC es la que muestra un mejor comportamiento. Además, este buen comportamiento se mantiene bastante estable para las diferentes dimensiones. La función TF-IDF es la siguiente mejor, seguida de la combinación analítica de criterios ACC. Las que muestran un comportamiento menos estable son WIDF y BinIDF. Cabe destacar que con esta reducción, las dimensiones de las representaciones B-IDF y WIDF resultan mucho mayores que para el resto de representaciones, dado un mismo valor de N .

Con la reducción Min-Max también las funciones FCC y TF-IDF, por ese orden, son las que ofrecen un mejor comportamiento. Sin embargo, en este caso, la función TF es la tercera mejor seguida de ACC. Cabe destacar que en ambos tipos de reducción los peores resultados se obtienen con la función WIDF.

A la luz de estos resultados, la combinación borrosa de criterios FCC, en la que se tiene en cuenta la intención del autor, traducida en el uso de etiquetas de enfatizado, mejora a las representaciones que no tienen en cuenta ese criterio. También mejora a la función ACC, que también lo contempla, pero mediante una combinación analítica.

Un aspecto destacable es que la función FCC no hace uso de información de la colección, sólo de la propia página, aspecto que consideramos muy importante en el contexto de las páginas web.

6. Conclusiones

En este artículo se han evaluado los resultados de realizar el *clustering* de una colección de referencia de páginas web, utilizando siete funciones de peso distintas y dos tipos de reducción del número de rasgos. El algoritmo de *clustering* utilizado ha sido un algoritmo de partición de la librería CLUTO. La calidad de las representaciones se ha medido utilizando la medida-F, comparando los resultados del algoritmo con los proporcionados en la colección de referencia.

Los mejores resultados se han obtenido utilizando la función de peso que realiza una combinación borrosa de criterios, FCC. Esos criterios son: (1) la frecuencia de un rasgo en la página, (2) si el rasgo aparece en el título, (3) si está enfatizado y, por último, (4) su posición en la página. Los criterios (2) y (3) se obtienen de las anotaciones HTML de las páginas web. La función que combina los mismos criterios, pero de forma analítica, ACC, obtiene peores resultados. Claramente el módulo borroso permite un mejor aprovechamiento del conocimiento que aportan los diversos criterios. Cabe destacar, además, que el buen comportamiento de la función FCC se mantiene estable aunque varíe la dimensión de la representación.

La función TF-IDF es la que presenta el segundo mejor comportamiento. Sin embargo, la función FCC tiene la ventaja adicional de que para calcularla sólo se necesita información de la propia página, no de la colec-

ción. En un contexto en el que una colección puede tener una tasa alta de actividad consideramos que éste es un aspecto que añade valía a la función.

La principal aportación de este trabajo ha sido probar que, usando funciones de peso basadas en combinaciones heurísticas de criterios, el *clustering* de páginas web se puede mejorar frente al uso de funciones basadas únicamente en frecuencias, tanto en el propio texto de la página como en la colección. Por último, estos resultados sugieren que el uso de conocimiento, a través de las anotaciones HTML, sobre la intención del autor de enfatizar cierto contenido del texto, puede ayudar a mejorar los resultados del *clustering* de páginas web.

Bibliografía

- Asirvatham, Arul Prakash y Kranthi Kumar Ravi. 2001. Web page classification based on document structure.
- Barfouroush, A. Abdollahzadeh, H.R. Motahary Nezhad, M. L. Anderson, y D. Perlis. 2002. Information retrieval on the world wide web and active logic: A survey and problem definition.
- Brin, Sergey y Lawrence Page. 1998. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30(1-7):107-117.
- Chakrabarti, S. 2002. *Mining the Web: Discovering Knowledge from Hypertext Data*. Morgan-Kaufmann Publishers.
- Fairthorne, R.A. 1961. The mathematics of classification. *Towards Information Retrieval*. Butterworths, London, páginas 1-10.
- Fresno, V. y A. Ribeiro. 2004. An Analytical Approach to Concept Extraction in HTML Environments. *Journal of Intelligent Information Systems - JIIS*, 22(3):215-235.
- Getoor, L. 2003. Link mining: a new data mining challenge. *SIGKDD Explor. Newsl.*, 5(1):84-89.
- Glover, Eric J., Kostas Tsioutsoulis, Steve Lawrence, David M. Pennock, y Gary W. Flake. 2002. Using web structure for classifying and describing web pages. En *WWW '02: Proceedings of the eleventh international conference on World Wide Web*, páginas 562-569. ACM Press.

- Gonzalo, J. 2004. Hay vida después de google? Informe técnico, In the Software and Computing System seminars. Escuela Superior de Ciencias Experimentales y Tecnología. Universidad Rey Juan Carlos. (<http://sensei.lsi.uned.es/> julio/).
- Good, I.J. 1958. Speculations Concerning Information Retrieval. *Research Report PC-78, IBM Research Center, Yorktown Heights, New York*.
- Guglielmo, E. J. y N. Rowe. 1996. Natural-language retrieval of images based on descriptive captions. *ACM Transactions on Information Systems*, 14(3):237–267.
- Harmadas, V., M. Sanderson, y M. D. Dunlop. 1997. Images retrieval by hypertext links. En *Proceeding of SIGIR-97, 20th ACM International Conference on Research and Development in Information Retrieval*, páginas 296–303.
- Iavernaro, F. 2004. Web Usage Mining Using Artificial Ant Colony Clustering and Genetic Programming. <http://www.knowledgeboard.com/cgi-bin/item.cgi?id=129238&d=pnd>.
- Jiayun, G. 2004. *Integrating Automatic Document Clustering into Web Log Association Rule Mining*. Ph.D. tesis, Faculty of Computer Science Dalhousie University, Canada.
- Karypis, G. 2002. CLUTO: A Clustering Toolkit. Informe Técnico 02-017, University of Minnesota, Department of Computer Science, Minneapolis, MN 55455.
- Lu, Qing y Lise Getoor. 2003. Link-based text classification. En *Proceedings of the IJCAI Workshop on Text Mining and Link Analysis*.
- Luhn, H. P. 1957. A statistical approach to mechanized encoding and searching of literary information. *IBM Journal of Research and Development*, 1(4):307–319.
- Merkel, D. 1998. *Text data mining. A handbook of Natural Languages Processing Techniques and Applications for the Processing of Languages as Text*. R. Dale and H. Moisl and H. Sommer (Eds). New York: Marcel Dekker.
- Needham, R.M. 1961. *Research on information retrieval, classification and grouping*. Ph.D. tesis, University of Cambridge; Cambridge Language Research Unit, Report M.L. 149.
- Pierre, J. M. 2001. On the Automated Classification of Web Sites. En *Linköping Electronic Articles in Computer and Information Science*, volumen 6.
- Ribeiro, A., V. Fresno, M. García-Alegre, y D. Guinea. 2002. A fuzzy system for the web page representation. *Intelligent exploration of the web. Springer-Verlag Group*, páginas 19–38.
- Rijsbergen, C. J. 1974. Foundations of evaluation. *Journal of Documentation*, 30:365–373.
- Salton, G. 1988. *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*. Addison-Wesley.
- Salton, G. y M. McGill. 1983. *Introduction to Modern information Retrieval*. McGraw Hill, New York.
- Salton, G. y C.S. Yang. 1973. On the specification of term values in automatic indexing. *Journal of Documentation*, 29(4):351–372.
- Sebastiani, F. 2002. Machine Learning in Automated Text Categorization. *ACM Computing Surveys*, 34(1):1–47.
- Sinka, M.P. y D. W. Corne. 2002. A Large Benchmark Dataset for Web Document Clustering. *Soft Computing Systems: Design, Management and Applications, Frontiers in Artificial Intelligence and Applications*, 87:881–890.
- Srihari, R. K. 1995. Automatic indexing and content-based retrieval of captioned images. *Computer*, 28(9):49–56.
- Zhongmei, Y. y C. Ben. 2003. Bidirectional Hierarchical Clustering for Web Mining. *IEEE/WIC International Conference on Web Intelligence (WI'03)*.

Reducción PF - CLUTO (k=10)

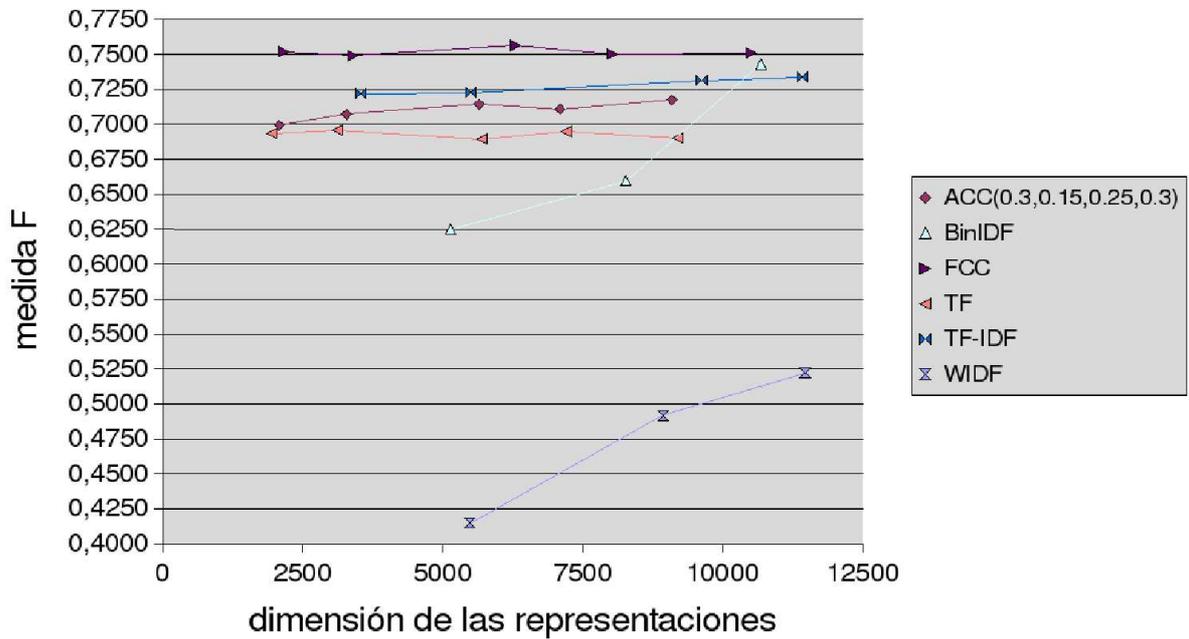


Figura 1: Resultados de *clustering* usando las propias funciones de peso como método de reducción de rasgos

Reducción MinMax - CLUTO (k=10)

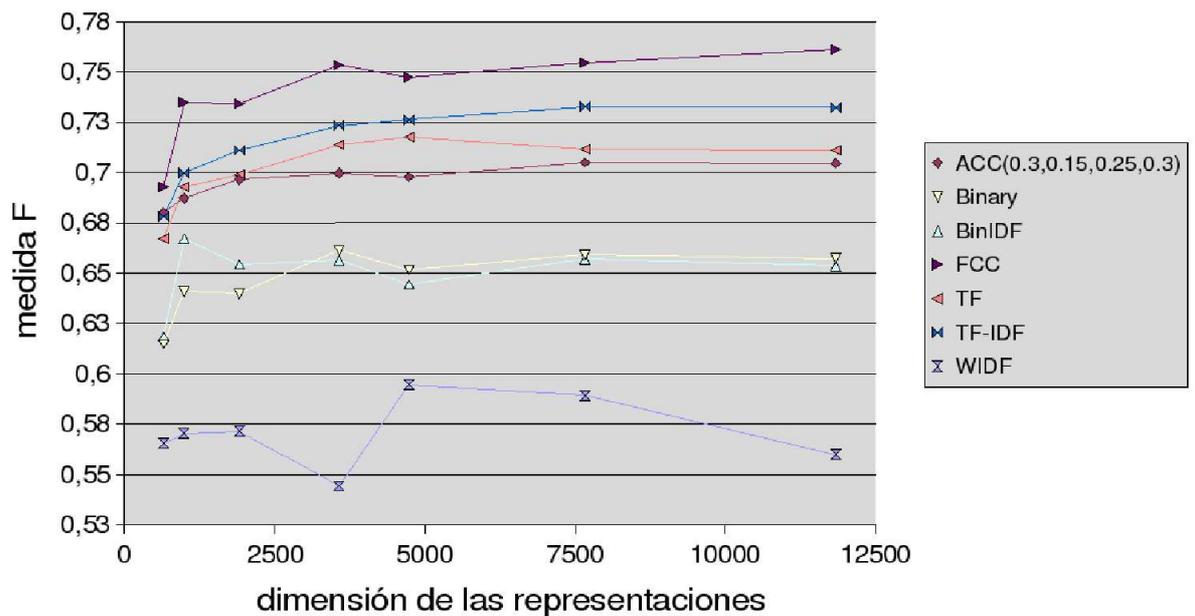


Figura 2: Resultados de *clustering* con reducción Min-Max