New Similarity Function for Scientific Articles Clustering based on the Bibliographic References

Lisvandy Amador Penichet [1], Damny Magdaleno Guevara [2],
Maria Magdalena García Lorenzo [2]

[1] Universidad Central "Marta Abreu" de Las Villas Instituto de Biotecnología de las Plantas,
Cuba

[2] Universidad Central "Marta Abreu de Las Villas", Departamento de Computación,
Cuba

lisvandy@ibp.co.cu, {mmgarcia, dmg}@uclv.edu.cu

Abstract. The amount of scientific information available on the Internet, corporate intranets, and other media is growing rapidly. Managing knowledge from the information that can be found in scientific publications is essential for any researcher. The management of scientific information is increasingly more complex and challenging, since documents collections are generally heterogeneous, large, diverse and dynamic. Overcoming these challenges is essential to give to the scientists the best conditions to manage the time required to process scientific information. In this work, we implemented a new similarity's function for scientific articles' clustering in based on the information provided by the references of the articles. The use of this function contributes significantly to discover relevant knowledge from scientific literature.

Keywords. Scientific paper; similarity function; clustering.

1 Introduction

The large number of existing scientific publications makes it difficult for users to identify relevant information from the results given by search engines [1]. The management of scientific information becomes increasingly more complex and challenging, mostly because the collections of documents are generally heterogeneous, large, diverse and dynamic. The automatic document clustering offers a possible solution to the problem of information overload, whereby users can quickly view the search results, using articles groups tagged, which have been grouped into categories of topics and sub-topics [1]. The specialized clustering of scientific articles has become a topic of particular interest, that why the authors have developed some works that proposed different alternative to resolve this problem.

Many of the papers found in the literature related to the scientific articles' clustering are aimed at using the co-citation index of articles to determine how similar they are [2-7]. Co-citation can be defined as the frequency which two articles are cited together by a new article [2, 4].

One of the earliest papers reported in the literature focusing specifically on scientific articles classification is the proposed by Garfield [6]. The developed method determines the relationship between the different pairs of articles, taking into account the number of times they are co-cited. For this, a grouping process is performed where a pair of articles belong to the same group if their co-citation's number exceeds a certain threshold. To classify a new document, the references are compared with the other articles' references of each cluster. The new document is going to be labeled with the header labels of those clusters which their references were matched.

In another approach that cluster scientific articles, the authors used the co-citation frequency of 24 Chinese journals of librarianship and information science to discover the relationship between them [3]. The results obtained allow to group the magazines into four fundamental groups, thus managing to relate those journals that deal with more related topics.

In [5], the authors propose different methods to improve the quality of the scientific articles'

Computación y Sistemas, Vol. 22, No. 1, 2018, pp. 93–102
ISSN 1405-5546
doi: 10.13053/CyS-22-1-2763

clustering based on the co-citation of the same ones. These methods, using different approaches, analyze the position where the co-citation appear in the text. The results of the applied experiments show that two references that are quite close in one article are more similar than two that are more distant. Also, some works have been developed using others approaches for the scientific papers clustering. Example of this is the co-citation analysis [1, 8-10].

In [8], the authors developed an algorithm to recommend scientific articles for users of an online community. According to authors, this approach manages to combine the merits of traditional collaborative filtering and probabilistic modeling.

In [9], the author presents a novel approach of monitoring to the problem of the multi-document recapitulation of scientific articles.

In [11], the authors show the results of a study on the automatic clustering, applied to scientific articles and journalistic texts in Brazilian Portuguese. One of the most recent works related to this subject is presented in [12], where a new methodology of scientific articles' clustering in semistructured format is developed. This methodology makes use of both, the structure and the content of the document, to achieve better results in the clustering.

To do this, the author develops a similarity function that allows to mix the results of the scientific articles' clustering by viewing each structural unit independently of each other and considering the article completely without taking into account the units for which it is composed.

The results obtained through the experiments shows that jointly exploiting the structure and content of the scientific articles considerably improves the clustering´s result. If this structure is also correctly exploited, in order to identify which of these parts are most significant when it is desired to know how similar two articles are, it is possible to increase the efficiency in the scientific articles' clustering, since it reduces the computational time by not having to process the entire document. At the same time, efficiency can be increased, because the extraction of terms focuses on parts of the article that provide more detailed and accurate information. For the aforementioned, it is proposed as a general objective of this work: to develop a similarity function for scientific articles based only on the information provided by the bibliographic references.

So, this paper presents: a new similarity measure that facilitates evaluating the degree of relationship between scientific articles based on the references.

2 Similarity Function Specializing in References

Scientific papers have specific characteristics that distinguish them from other documents, including the selection of key words in the document and the presence of bibliographic references. If these distinctive features are used in terms of achieving a better scientific papers' clustering, relevant results can be obtained.

2.1 Similarity Function for Scientific Articles Clustering

The proposed method uses as input the result of an information retrieval process [13]. The output is homogeneous clusters of related documents and the quality of the clustering; guaranteeing control for the evaluation of the results. Figure 1, shows a graphical view of the new form of clustering based on the author, title and year subunits, all belonging to the bibliographic references.

2.1.1 Representation of the Corpus Obtained

By working with the title subunit, the VSM representation was selected and a change to represent the author subunit. No need trying every word that makes up the name of an author as an independent term, this could make some discrepancy while verifying how similar two documents are. That why the content of the author subunit will be treated as a text string, and is considered the full name of an author as a single term.

2.1.2 Removing Subunit Terms Title

To obtain the representation of the title subunit, it initiates with a sequence of tokens and a sequence of indexed terms based on these is produced.

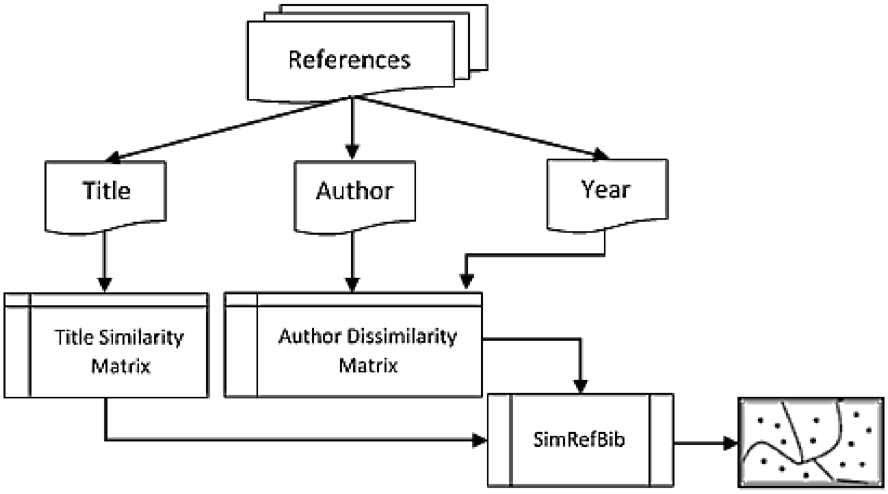

Fig. 1. Graphical representation of clustering method proposed

The next step is to select only those tokens that are relevant words in the title subunit of the analyzed document. In this step, it is considered a candidate word as relevant when its frequency exceeds the *afw* threshold; *afw* is variable and depends of references' number (*RBN (i))* in the document analyzed, so:

$$afwi = \begin{cases} 2 & if \quad RBNi \leq 10, \\ 3 & if \quad 10 < RBNi \leq 20, \\ 4 & if \quad 20 < RBNi \leq 25, \\ 5 & i.o.c. \end{cases} \tag{1}$$

After obtaining the relevant tokens of title subunit, the process of joining tokens is required. This process is important because the tokens obtained cannot be seen as one-off simple terms.

The first step of joining process token consists of finding the frequency of the relevant tokens (taken in pairs) in the title subunit of the references.

After obtaining the relevant tokens' pairs, which would be those which exceed the *afw* threshold, we analyze whether some of the formed pairs can be joined, this is done only for pairs that the initial substring of the first pair matches the final substring of second pairs, or vice versa, taking as the initial substring, the first word of the pair, and as final substring, the last word.

After finishing the process of joining tokens, the relevant phrases for the document are obtained, as well as the importance of each phrase. Definition 1.1 denotes the importance of the relevant word *i* in document *k*.

Definition 1.1 (Importance of Relevant Word):

Be *fki* the frequency of the relevant word *i* in document *k, RBN(i)* the number of references in the document *i*, the importance of word *i* in document *k* is defined as:

$$Impk,i = \begin{cases} fki/RBNi & if \ fki/RBNi < 1, \\ 1 & i.o.c. \end{cases} \tag{2}$$

After the joining process, there are some one-off or independent tokens yet, which are not relevant words, only those with a frequency of greater than or equal to 25% of the references' number of the document, or absolute frequency of occurrence in the is references greater than 10 are considered relevant.

Nevertheless, some of the tokens that were joined with others prevail as a separate token, if they exceed the threshold of 25% of the references or their frequency is greater than 10, these tokens also become part of the relevant words.

The process of joining tokens will not be applied in the documents that don't have terms that exceed the *afw* threshold, due to that fact that if as an independent term it does not exceed the threshold, it obviously will not exceed as attached terms.

Distinguishing Relevant Words:

When the relevant words for each document are obtained, it can be seen that there are some of these words that recur significantly in the collection. These words are called distinguishing relevant words because as they are present in a considerable number of documents, they provide a greater degree of similarity and can determine more precisely related documents clusters.

The selection process of the distinguishing relevant words is described in the following steps:

1. Calculate the occurrences' number of each relevant word.
2. Sort the downward words according to the number of times they appear.
3. If k = 0 go to step 5, where k is the number of distinct relevant words that the user decides to select.
4. Select the k first words. Exit.
5. Select all the words that their frequency in the collection is greater than 5. Exit.

The processing of the title subunit ends with the distinguishing relevant words extraction, once finished this process it can proceed to calculate the Title Similarity (*TS*), between the documents.

2.1.3 Removing Terms of the Author Subunit

For the representation of the author subunit a modification of VSM is used, considering the author as a single term and storing its importance (which would be the number of times it is referenced in the document), the interval of years that this author is referenced in the document in question is stored, and the number of times that it is not referenced as the main author is stored too.

It is essential in the processing of the author subunit to normalize the storing process of the authors' names.

Example:

Suppose we have the author Juan Pablo Pérez Rodríguez, which may appear referenced in the following ways (and even others): Juan P. Pérez Rodríguez, JP Perez, J. Perez, J. Perez, JP Perez.

To solve this problem, the author's name is standardized as follows: *XY Name*, where *X*, *Y* are the initials of the author's first name and *Name* will be the surname of the author.

Is important to distinguish some work around the disambiguation and normalization of the authors' names, such as by [14], in which a critical analysis of the main existing approaches in the literature to solve the problem of authors disambiguation in scientific publications is done. However, the most referenced work provides solutions that use the metadata of digital magazines or web as an information's source which would be difficult to adapt to our proposal.

Another problem that can be found in the References is *et.al.* term used to refer to an authors' group; if this term is found during lexical-graph analysis, it will not be saved.

2.2 Calculating the Similarity Between Scientific Articles

Title and author subunits are treated separately in the computation of similarity. The Year subunit is used according the author subunit. That is because two articles that have similar years in the references do not have why to approach the same subject.

2.2.1 Calculating the Author Dissimilarity

For the computation of the dissimilarity between the documents, considering author subunit, it used *DisAut* measure which is defined as follows:

Definition 1.2 (*DisAut*): Given the documents *i* and *j*, the dissimilarity measure *DisAut (i, j),* is defined to indicate how different is this pair considering authors referenced therein; this measure is formalized mathematically in Equation 3:

$$DisAut i,j = WD(i,j)^{1-BS(i,j)} \times \, UD i,j^{BSi,j}. \qquad (3)$$

BS (i, j) indicates the binary similarity between the pair of documents analyzed, *WD* (*i, j*), is the weighted dissimilarity between them and *UD (i, j),* is the unweighted dissimilarity.

The general idea of computing the binary similarity (*BS* (*i, j*)), is based on the hypothesis that, if two documents refer in a high percentage to the same authors, these documents must treat similar subjects, then dissimilitude without weight would then apply.

The unweighted dissimilarity does not consider the range of years where each author is referenced. It is probable that if an author is referenced in two documents d_i and d_j but the difference between the reference intervals of each one of the documents is considerably large, this author would treat different issues, or what is the same, they may have changed his research line, but this probability is reduced to the extent that increases the number of same authors are referenced in both scientific articles, because it would be very coincidental that several authors change together its research line.

Definition 1.3 (Binary Similarity): Given the documents *i* and *j*, binary similarity between them is defined as:

$$BSi,j = \begin{cases} 1 & if \quad STFi,j \geq \varepsilon, \\ 0 & i.o.c. \end{cases} \tag{4}$$

Being ε the similarity threshold (recommended ε = 0.5). *STF (i, j),* is defined by equation 5:

$$STF(i,j) = \frac{2 \times \sum_{k=0}^{n} STFPi_k, j_k}{\sum_{k=0}^{n} NCi_k, j_k}. \tag{5}$$

In the above equation *n* indicates the number of authors referenced in the collection and *STFP* (i_k, j_k), and *NC* (i_k, j_k), are defined as follows:

$$STFPi,j = \begin{cases} 1 & if \quad Ci_k \neq 0 \; y \; Cj_k \neq 0, \\ 0 & i.o.c., \end{cases} \tag{6}$$

$$NCi,j = \begin{cases} 1 & if \; C_{ik} - Cnp_{ik} \neq 0 \; o \; C_{jk} - \\ & Cnp_{jk} \neq 0 \;\; o \; STFPi,j \neq 0, \\ 0 & i.o.c. \end{cases} \tag{7}$$

C_{ik} and C_{jk} indicate the number of times in which the author *k* is referenced in the documents *i* and *j* respectively; Cnp_{ik} and Cnp_{jk} indicate the number of times that the author *k* is referenced as non-main author in these documents.

The weighted dissimilarity *WD* indicates the difference between a pair of documents taking into account: the authors referenced, the number of times they are referenced and the range of years that each author is referenced. The mathematical formalization of *WD* appears in the equation 8.

The weight is given by the operator wk which varies depending on the range of years in which the author k is referenced in the documents i and j. Thus, if years intervals intersect or are close (considered a close intervals pair as a neighborhood of 5 years), would apply the difference arithmetic operator (-), to the number of references of the author k. Otherwise, the addition arithmetic operator (+), applies:

$$WD(i,j) = \frac{\sum_{k=0}^{n} \left| C_{ik} w_k C_{jk} \right| - AR_i + AR_j}{CA_i - AR_i + CA_j - AR_j}, \tag{8}$$

where *n* indicates the number of authors referenced in the documents collection.

C_{ik} and C_{jk} represent the number of times that the author *k* is referenced in documents *i* and *j* respectively.

CA_i and CA_j represent the sum of references of all authors in *i* and *j* respectively.

AR_i and AR_j represent the sum of the times that the authors of the documents *i* and *j* provide noise.

An author *k* provides noise to find the dissimilarity between a pair of objects *i* and *j*, (with $C_{ik} \geq C_{jk}$), if the number of times in which he is referenced as non-main author on the paper *i* is greater than zero and the difference between the number of total references (C_{ik}, C_{jk}), for this author in document *i* and document *j* is greater than zero too. The noise value (*NV (k, i)*), provided by the author is defined as:

$$NVk,i = \begin{cases} Cnp_{ik} & if \quad (C_{ik} - C_{jk} - Cnp_{ik}) \geq 0, \\ C_{ik} - C_{jk} & i.o.c. \end{cases} \tag{9}$$

Cnp_{ik} represents the number of times that the author *k* is referenced as non-main author on the paper *i*.

The unweighted dissimilarity *UD* just varies from dissimilarity *WD* in the w_k operator, which always is applied as difference operator (-). *UD* is defined mathematically in Equation 10:

$$UD(i,j) = \frac{\sum_{k=0}^{n} \left| C_{ik} - C_{jk} \right| - \left(AR_i + AR_j \right)}{CA_i - AR_i + \left(CA_j - AR_j \right)}. \tag{10}$$

2.2.2 Calculating the Similarity Degree (ST)

The title similarity between a pair of documents *i* and *j* is defined by the degree of similarity between the relevant words of the title subunit obtained for each document, so the first step will be to calculate the values of similarity between each pair of words.

Suppose that we have the document d_i with the relevant words (PI_1, PI_2, ..., PI_n), and the document d_j with relevant words (PJ_1, PJ_2, ..., PJm). The *Sim_matrix*(*nxm*) matrix is formed, where *n* is the number of relevant words of d_i and *m* the number of relevant words d_j.:

$$Sim\_matrix(i,j) = 1 - JWPI_i, PJ_j, \qquad (11)$$

where *JW*(*i*, *j*), is the Jaro Winkler distance [15].

```
Input: Two-dimension array DOC1 and DOC2
with N*2 and M*2 elements respectively. N y
M number of relevant words of DOC1 and
DOC2, DOC1[i,1]: relevant word, DOC1[i,2]:
word's relevance. PRS vector with length k,
where PRS[i]= Significant Relevant Word.

Output:sim=$WeigthSumN/(\max(N,M) - MatchWords + WeigthSumD)$,
Similarity between documents di, dj)

Begin

1. Sim_matrix □ Similarity Matrix compute,
   according equation 11.
   For each PRSi do
     if DOC1[s1,1] and DOC2[s2,1]
     contains PRSi
    then
         PRICList[j,1] □ s1,
         PRICList [j,2] □ s2
   End_For

2. minCPR □ min(N,M)

3. Repeat
     $maxV$ □ $maxSim\_matrix[i,j] \mid i \notin I_{selected} \wedge j \notin J_{selected}$
     if max ≥ 0.9 o PRICList contains (i, j)
  then
   increase $MatchWords$
   increase t
    $I_{selected} \leftarrow I_{selected} \cup i$
   $J_{selected} \leftarrow J_{selected} \cup j$
    $WeigthSumN \mathrel{+}= \sqrt{maxV}\,\frac{(DOC[i,2]+DOC[j,2])}{2}$
           $WeigthSumD \mathrel{+}= \max(DOC[i,2], DOC[j,2])$
   else
     Sim_matrix[i,j] □0
 until (t ≥ minCPR) o ($maxV$=0)

end
```

Algorithm 1. Algorithm to calculate the Title Similarity (*TS*) between two scientific papers

The computation of the similarity degree is formalized in Algorithm 1. In general, this algorithm searches in the similarity matrix *Sim_matrix* the word's pair (*i*, *j*), with maximum similarity value, if this value exceeds the threshold established (for this case 0.9), or both words containing one distinguishing relevant word, the similarity value of these words is considered in the function and further multiplied by the average of the weights of these words, the weight of a word *k* in a document *i* is calculated by expression 2.

It looks at most *q* pairs of words, *q* is the minimum between *n* and *m*. The overall similarity value obtained shall be standardized with the sum of: the difference of the maximum value between *m* and *n* with respect to the number of words match, and the sum of the maximum weights for each pair (*i*,*j*), that was selected.

2.2.3 General Measure of Similarity

For the calculation of the overall similarity it is necessary to mix the value obtained for Title Similarity between documents and Author Dissimilarity value of them. Clearly, these values do not have equal weight because it is not possible to relate an unambiguous manner an author with a given topic, and the relevant words certainly determines specific issues.

Therefore, it is used in this work as value of greater weight to measure the similarity between a pair of documents the Title Similarity obtained. The value obtained for the Author Dissimilarity will be used as a positive influence on the overall similarity.

Mathematically we formalize the general similarity from the references (*SimRefBib (i,j)*), between two documents *i* and *j* as follows:

$$SimRefBibi,j = \begin{cases} TSi,j^{DisAuti,j} & if\; TSi,j > 0, \\ 1 - DisAuti,j & if\; TSi,j = 0. \end{cases} \qquad (12)$$

3 Validation

In the CEI-UCLV there is a large number of scientific articles and documents related to various

Table 1. Description of case studies

No. Corpus	Number of documents	Number of lasses	Themes it treats
Documents' sets in XML format made from retrieved documents from the ICT site of Centro de Investigaciones en Informática de la Universidad Central "Marta Abreu" de Las Villas http://ict.cei.uclv.edu.cu			
1	32	2	Fuzzy Logic, SVM
2	25	2	RST, Association Rules
3	32	2	RST, SVM
4	28	2	Association Rules, Fuzzy Logic
5	32	2	Association Rules, SVM
Documents collection from the IDE-Alliance repository, internationally used to evaluate grouping. Provided by the University of Granada. Spain.			
6	28	3	Copula, CL, Belief Propagation
7	19	2	Copula, Belief Propagation
Documents belonging to the ICT site and the IDE-Alliance repository			
8	41	4	Belief Propagation, RST, Copula, CL
9	29	2	Copula, SVM
10	38	3	Belief Propagation, Copula, SVM

topics of research, available to the network of the Ministry of Higher Education (MES).

The first case study was formed from files of the site (ICT), to verify the benefits of the new function of similarity in the information retrieval and extraction of knowledge that the users are seeking. The second case study is a compilation of documents from the repository IDE-Alliance, internationally used to evaluate clustering which is provided by the University of Granada, Spain.

The third case study constitutes a set of corpora formed from the union of documents belonging to the two case studies mentioned above.

All data sets have one objective feature, therefore there is the reference classification for each of them, and specifically in the first study case this feature was obtained based on the criterion of experts. The remaining collections were acquired with the reference classification. In Table 1, we offer a description of the case studies.

The first experiment consisted of verifying if there were differences when we apply the similarity's function to the three case studies mentioned above with respect to other functions existing in the literature. *Jaccard's* functions [16], *Cosine* and *Dice* [17], were selected because they report good results in the clustering of text.

The clustering algorithm proposed in INEX by [18], was selected. Based on the results of this algorithm the Overall F-Measure (OFM), Micro-Purity and Macro-Purity measures were applied to establish a comparison between the different functions listed above.

To apply the non-parametric Friedman [19], test for Micro-Purity measure significance values lower than 0.05 are observed which indicates that there are significant differences between the compared populations. In Figure 2, it is observed that differences between the mean values of these samples.

When applying the test of Nemenyi (Figure 3), was obtained that there are significant differences between the *SimRefBib* and *Cosine* functions and the *SimRefBib* functions and *Dice* for $q = 0.05$. For

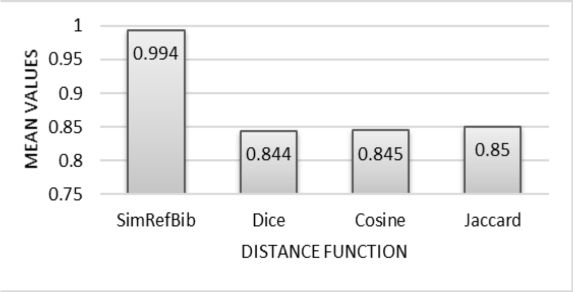

Fig. 2. Mean values obtained for Micro Purity measure in each distance applied

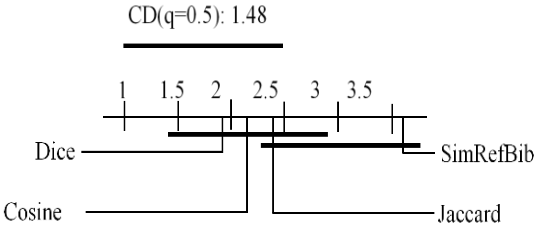

Fig. 3. Test of Nemenyi with q = 0.05 for Micro-Purity values obtained by applying the algorithm K-Star to each of the functions

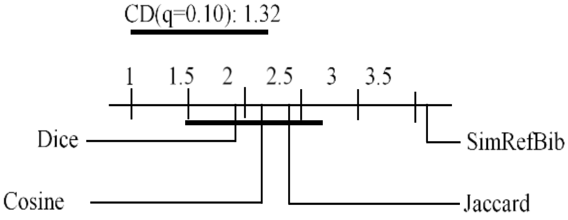

Fig. 4. Test of Nemenyi with q = 0.10 Micro-Purity values obtained by applying the algorithm for K-Star to each of the functions

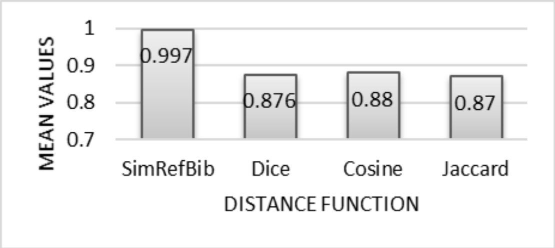

Fig. 5. Mean values obtained for Macro Purity measure in each distance applied

q = 0.10 we obtain significant differences comparing *SimRefBib* function to the rest of the functions as shown in Figure 4. These differences always achieved better results for *SimRefBib* than for the remaining function.

Also, non-parametric test of Friedman for Macro-Purity measurement is applied. In Figure 5 shows the differences among mean values obtained.

The test of Friedman threw as a result that there are significant differences between the compared samples. Nemenyi test was applied (Figure 6), and it obtained that for *q* = 0.05, there are significant differences between the *SimRefBib* function and *Dice* and *Jaccard* functions; for *q* = 0.10 there were significant differences between the *SimRefBib* function and the rest of the functions, see Figure 7. Significant differences have always had a behavior in favor of the function proposed in this research.

Non-parametric Friedman test showed that there are no significant differences between the OFM values obtained for each of the tested functions. However, we came to the conclusion that the function proposed in this research have a behavior more stable than the rest of the functions tested as shown in Figure 8.

The second experiment consisted of verifying if there are differences when applying the similarity's function to the three case studies mentioned above with respect to apply the *OverallsimSUX* similarity function [20].

This last function makes use of all the structural units of the scientific article to obtain the similarity matrix. That is because obtaining the groups of related documents when applying this function has a higher computational cost than when applying the SimRefBib function that only makes use of bibliographical references.

To apply the non-parametric Wilcoxon test for Micro-Purity, Macro-Purity and OFM measures significance values lower than 0.05 are observed which indicates there are significant differences between the compared populations. In Figure 9, differences between the mean values of these samples are observed.

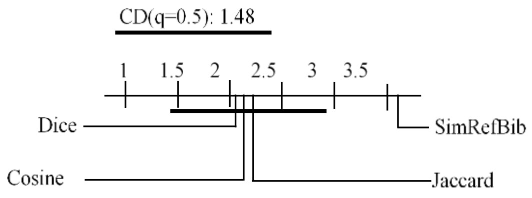

Fig. 6. Nemenyi Test with *q* = 0.05 for Macro-Purity values obtained by applying the algorithm K-Star to each of the functions

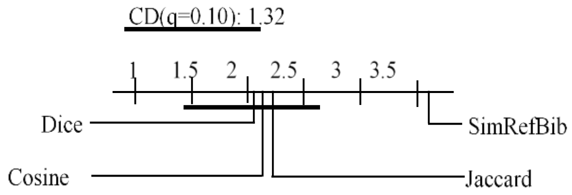

Fig. 7. Nemenyi Test with *q* = 0.10 for Macro-Purity values obtained by applying the algorithm K-Star to each of the functions

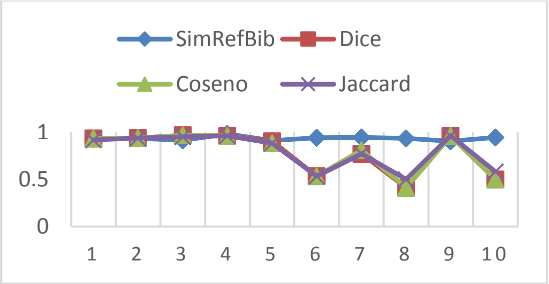

Fig. 8. Behavior of the OFM measure for each of the analyzed functions

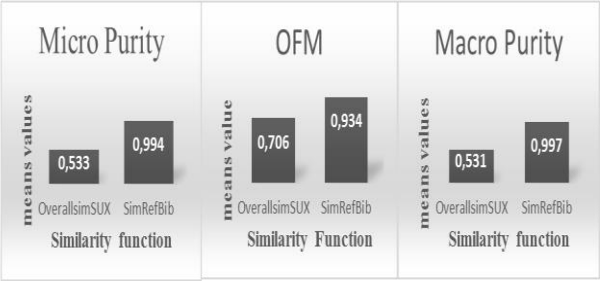

Fig. 9. Differences between the mean values of analyzed samples conclusions

4 Conclusions

There is a growing conceptual theoretical base on documents' clustering in semi-structured format. However, the main works reported in the literature are focused on treating the documents in their entirety and not focus on the relevant parts of these, for example references in scientific articles.

The similarity function proposed for the documents' comparison captures the degree of similarity between the bibliographic references of documents, taking the relationship existing between the subunits present in a reference as genesis.

Evaluation through experiments and studies defined cases, using the K-Star clustering algorithm, yielded better results with the function proposal that other existing variant in the literature.

The comparison of the results obtained by applying the SimRefBib function and the OverallsimSUX function in the scientific articles' clustering show that it is much more feasible to focus on key parts of the article, such as bibliographical references, since it reduces processing time and increases effectiveness in the clustering.

References

1. **Aljaber, B., Stokes, N., Bailey, J., & Pei, J. (2010).** Document clustering of scientific texts using citation contexts. *Information Retrieval,* Vol. 13, No. 2, pp. 101–131. DOI: 10.1007/s10791-009-9108-x.
2. **Small, H. (1973).** Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the Association for Information Science and Technology,* Vol. 24, No. 4, pp. 265–269. DOI: 10.1002/asi.4630240406.
3. **Hu, C.P., Hu, J.M., Gao, Y., & Zhang, Y.K. (2010).** A journal co-citation analysis of library and information science in China. *Scientometrics,* Vol. 86, No. 3, pp. 657–670. DOI: 10.1007/s11192-010-0313-6.
4. **Wang, X., et. al. (2013).** Knowledge-transfer analysis based on co-citation clustering. *Scientometrics*, Vol. 97, No. 3, pp. 859–869. DOI:10.1007/s11192-013-1077-6.
5. **Boyack, K.W., Small, H., & Klavans, R. (2013).** Improving the accuracy of co-citation clustering

using full text. *Journal of the American Society for Information Science and Technology,* Vol. 64, No. 9, pp. 1759–1767. DOI: 10.1002/asi.22896.

6. **Garfield, E., Malin, M.V., & Small, H. (1975).** A system for automatic classification of scientific literature. *Journal of the Indian Institute of Science,* Vol. 57, No. 2, pp. 61.
7. **Small, H. & Sweeney, E. (1985).** Clustering the science citation index® using co-citations: I. A comparison of methods. *Scientometrics,* Vol. 7, No. 3-6, pp. 391–409. DOI: 10.1007/BF02017157.
8. **Wang, C. & Blei, D.M. (2011).** Collaborative topic modeling for recommending scientific articles. *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining,* ACM. DOI: 0.1145/2020408.2020480.
9. **Shankar, R. (2012).** *Evolutionary Document Clustering and Summarization of Scientific Articles using Frequent Item sets.* International Institute of Information Technology, Hyderabad.
10. **West, J.D., Wesley-Smith, I., & Bergstrom, C.T. (2016).** A recommendation system based on hierarchical clustering of an article-level citation network. *IEEE Transactions on Big Data,* Vol. 2, No. 2, pp. 113–123.
11. **Afonso, A.R. & Duque, C.G. (2014).** Automated text clustering of newspaper and scientific texts in brazilian portuguese: analysis and comparison of methods. *JISTEM-Journal of Information Systems and Technology Management,* Vol. 11, No. 2, pp. 415–436. DOI: 10.4301/S1807-17752014000200 011.
12. **Magdaleno, D. (2015).** *Metodología para el agrupamiento de documentos semiestructurados, in Ciencia de la Computación.* Universidad Central Marta Abreu de Las Villas, pp. 131.
13. **Berry, M.W. & Castellanos, M. (2004).** Survey of text mining. *Computing Reviews,* Vol. 45, No. 9, pp. 548.
14. **Alonso-Sierra, L.E., Hidalgo-Delgado, Y., & Leiva-Mederos**. **A.A. (2014).** Desambiguación del nombre de los autores en revistas científicas. *Revista Cubana de Ciencias Informáticas,* Vol. 8, No. 3, pp. 149–169.
15. **Winkler, W.E. (1988).** Using the EM algorithm for weight computation in the Fellegi-Sunter model of record linkage. *Proceedings of the Section on Survey Research Methods,* American Statistical Association.
16. **Lin, Y.S., Jiang, J.Y., & Lee, S.J. (2014).** A similarity measure for text classification and clustering. *IEEE transactions on knowledge and data engineering,* Vol. 26, No. 7, pp. 1575–1590. DOI: 10.1109/TKDE.2013.19.
17. **Vargas-Flores, S.I. (2016).** *Comparación de medidas de similitud para desambiguación del sentido de las palabras utilizando rankeo de grafos.* Universidad Autónoma del Estado de México, Ciudad México, pp. 125.
18. **Pinto, D. (2009).** BUAP: Performance of K-Star at the INEX’09 Clustering Task. *Focused Retrieval and Evaluation,* Springer, pp. 434–440.
19. **Friedman, M. (1937).** The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association.* Vol. 32, No. 200, pp. 675–701.
20. **Magdaleno, D., Fuentes, I.E., & García, M.M. (2015).** Clustering XML Documents Using Structure and Content based on a New Similarity Function OverallSimSUX. *Computación y Sistemas,* Vol. 19, No. 1, pp. 151–161. DOI:10.13053/CyS-19-1-1922.

Article received on 04/09/2016; accepted on 14/10/2016.
Corresponding author is Lisvandy Amador Peniche.