Clustering Abstracts instead of Full Texts*

Pavel Makagonov[1], Mikhail Alexandrov[2], and Alexander Gelbukh[2,3]

[1] Mixteca University of Technology, Mexico,
mpp2003@inbox.ru
[2] Center for Computing Research,
National Polytechnic Institute, 07738, DF, Mexico
dyner@cic.ipn.mx, gelbukh@gelbukh.com, www.Gelbukh.com
[3] Computer Science and Engineering Department,
Chung-Ang University, 156756, Seoul, Korea

Abstract. Accessibility of digital libraries and other web-based repositories has caused the illusion of accessibility of the full texts of scientific papers. However, in the majority of cases such an access (at least free access) is limited only to abstracts having no more then 50-100 words. Traditional keyword-based approach for clustering this type of documents gives unstable and imprecise results. We show that they can be easy improved with more adequate keyword selection and document similarity evaluation. We suggest simple procedures for this. We evaluate our approach on the data from two international conferences. One of our conclusions is the suggestion for the digital libraries and other repositories to provide document images of full texts of the papers along with their abstracts for open access via Internet.

1 Introduction

Frequently one has to cluster documents (e.g., scientific papers, patent applications, etc.) basing on short abstracts instead of full-text documents. A typical approach to document clustering in a given domain is to transform the textual documents to vector form basing on a list of keywords (linguistic indices) and to use well-known numerical procedures of cluster analysis [10]. The list of keywords is constructed from a training document set belonging to the same domain. However, with such an approach applied to abstracts we have:

- very unstable results with regard to slight changes of the keyword list or document set,
- very inexact results as compared to a human expert's opinion.

The former circumstance is due to extremely small size of documents, which leads to very small absolute frequencies of keywords. The reason of the latter

* Work done under partial support of Mexican Government (CONACyT, SNI, CGPI, COFAA) and Korean Government (KIPA professorship). The third author is currently on Sabbatical leave at Chung-Ang University.

P. Makagonov, M. Alexandrov, A. Gelbukh. Clustering Abstracts instead of Full Texts. Text, Speech and Dialogue (TSD-2004). Lecture Notes in Artificial Intelligence, N 3206, Springer-Verlag, 2004. pp. 129-135.

circumstance is the difference between the contents of abstracts and the papers: indeed, the abstracts explain the goals of the research while the paper explains the methods used.

Though there exists extensive literature on information retrieval [2, 12], the problem of clustering short documents is not well-studied. We are not aware of any comparison of clustering abstracts versus full-text papers, even if this is of special interest in the era of Internet. The only reports we are aware of concern categorization of short documents based on preliminary training [5, 7, 13]. However, this is a different situation, because we deal with clusters unknown beforehand rather than with predefined categories.

In this paper we suggest *simple modifications* of the traditional approach, which can significantly improve the results of clustering:

- For selecting keywords from the word frequency list, we consider objective criteria related to relative frequency of words with respect to general lexis and the expected number of clusters.
- For measuring similarity between documents, we use a weighted combination of cosine and polynomial measures.

2 Relationships between Documents

2.1 Constructing a Keyword List

We use the term *domain* to refer to a topic reflected in the whole document collection. A domain dictionary (DD) is a keyword list characterizing a specific domain (e.g., *chemistry*, *computational linguistics*, etc.). Such keywords are linguistic indices providing numerical representation of textual documents and the metric relations between them [12]. In the word frequency list all words having the same base meaning are joined and presented in the truncated (stemmed) form The algorithm uses empirical formulas for testing word similarity, which makes it almost language independent [9].

Given the word frequency list, we use a set of criteria [8] for filtering out stopwords: only those words W are included in the DD for which

1. $F_{Dom}(W) \gg F_{Com}(W)$; namely, $F_{Dom}(W)/F_{Com}(W) > k$, where $F_{Dom}(W)$ and $F_{Com}(W)$ are the frequencies of the word W in our document collection and in the general balanced corpus of the given language (common use), respectively, and
2. The relative number N of documents in which they occur is between two thresholds: $N_L < N < N_H$.

The parameter k is determined empirically. Its value is related to the statistical estimation of the mean error in the measuring of the frequencies due to a limited size of the sample texts. Namely, one or two occurrences of any low frequency word in a text doubles its frequency count. Because of the random

nature of these occurrences the error of the frequency estimation becomes comparative to the frequency itself. To avoid such a situation, a reasonable value for k must be greater than 3 or 4 for low frequency words in short texts.

The parameters N_H and N_L define how fine-grained the obtained classification is. Namely, they determine the maximum and minimum size of the expected clusters and consequently the minimum and maximum number of the clusters. To obtain 5–10 clusters, each word should occur in approximately 10% to 20% of the documents. Of course, this connection between the number of clusters and the number of documents is approximate and assumes a uniform distribution of the word by the documents. In practice (with non-uniform distribution), these boundaries should be at least doubled: to obtain 5–10 clusters, each word should occur in 5% to 40% of the documents.

We believe that these two criteria are more relevant to the task of clustering documents from a new domain or sub-domain than other statistical criteria. In particular, the criterion relying on the *tf-idf* measure [2] can not be used for abstracts because it does not work well when all words have very low frequency. In addition, it does not take into account the a priory information about the number of clusters.

2.2 Combined Measure of Document Similarity

Let x_{ij} be the number of occurrences of the keyword w_j in the text T_i normalized by its size M_i, where M_i is the number of running words in the document (excluding stop-words such as prepositions, etc.); such normalization reduces all estimations to the per word average. With this vector representation, the distance between two documents can be evaluated using the well-known cosine or polynomial (linear or quadratic) measures [10] and the combination between them:

$$
\begin{aligned}
D &= \alpha D_c + \beta D_p, \quad \alpha + \beta = 1, \\
D_c &= 1 - \frac{\sum\limits_k (x_{1\,k}, x_{2\,k})}{\|x_1\|\,\|x_2\|}, \\
D_p &= \sqrt[p]{\sum_k (x_{1\,k} - x_{2\,k})}, \quad p = 1, 2.
\end{aligned}
$$

It is important to note that the coefficients α and β do not reflect the real contribution of each measure to the combined one. In fact the density of keywords in a document does not exceed 5% for almost all cases interesting in practice. In the polynomial measure the density of 5% defines the distance of 0.05 between documents after normalization by the number of words. We assume that the maximum distance can reach even 0.1 for some specific collections, such as conference programs, résumés, etc. So we normalize the polynomial measure once more in 0.1 for all the documents under consideration. The combined measure

was introduced in our paper [1] as one of the parameters used in our Document Investigator toolset, though we did not discuss there its possible applications.

Our hypothesis is that in case of clustering the abstracts such a measure can improve the accuracy of automatic clustering as compared with the expert opinions. Indeed, abstracts communicate first of all the goals of a paper but not the methods used. In this case the combined measure may give better results than the pure cosine or the pure polynomial measure, because the former one evaluates the closeness of the proposed methods by the closeness of the themes of the abstracts (which due to the mentioned difference leads to inexact results), while the latter one overemphasizes domain representativity (which due to low keyword density leads to unstable results). The experiments described in the next section support this hypothesis.

3 Experiments with Web-Retrieved Documents

3.1 Data and Methods Used

In all our experiments, we compare the results of *automatic* clustering of *abstracts* and *manual* clustering of *full-text* papers. The latter is considered as the ideal solution. The goal of our experiments is to investigate the dependence of the results on: the parameters of the combined measure, the clustering methods, the domain dictionaries, the broadness of the domain, and the type (papers and abstracts) of the documents.

Experiments with the combined measure With different parameters of the combined measure $D = \alpha D_c + \beta D_p$ we obtained different clusters. We experimened with different clustering methods, domain dictionaries, and data. We tried the following combinations of parameters:

Table 1. Parameters of combined measure we tried

$\boldsymbol{\alpha}$	1	1	1	1	1	0.5	0.5	0	0
$\boldsymbol{\beta}$	0	0.5	0.5	1	1	1	1	1	1
$\boldsymbol{p}$	–	1	2	1	2	1	2	1	2

In Table1, α and β are the weights of cosine and polynomial measure and p stands for the power of the polynomial measure. The coefficients α and β are given before normalization to 1; $p = 1$ or 2. Note that the contribution of the polynomial measure increases from left to right in the table.

In our experiments reflected in the Table 2 we did not look for the best combination of the parameters, i.e. the one providing the best coincidence between automatic and manual clustering. Instead, wee were only interested in sensibility of the clustering results to the parameters of the combined measure.

Experiments with different methods of clustering There are more methods and their modifications used in cluster analysis than authors working in this area. Extensive literature is devoted to such methods and their applications in text processing [10].

For simplicity, we tried in our experiments only two methods: the simplest hierarchical method (*nearest neighbor*) and the simplest non-hierarchical method (*K-means*) [4]. The former method builds a dendrite and then eliminates the weak connections so that instead of one tree several sub-trees appear. Each subtree is considered a cluster. In the latter method the desired number of clusters is defined by the user.

These two methods are the simplest and in a certain sense the most different from each other, i.e., they give the least coincident results on various data sets as compared with other pairs of clustering methods [11]. So, the coincidence of their results would be a strong indication of stability of obtained clusters.

Experiments with domain dictionaries In our experiments we considered two sets of abstracts. For each of them we constructed two dictionaries using the following parameters of keyword selection: for one set, $k = 4$, $N_H = 40\%$, $N_L = 5\%$, and for the other set, $k = 7$, $N_H = 70\%$, $N_L = 5\%$. The reason for such a selection of the values is the following:

- If $k < 4$ then the results prove to be very sensitive to low frequency words, while with $k > 7$ the dictionaries prove to be too small, which causes problems in clustering;
- The pairs $N_H = 40\%$, $N_L = 5\%$ and $N_H = 70\%$, $N_L = 5\%$ correspond to the expected number of clusters of 5–10 and 2–10, respectively, which are adequate for one-level clustering.

With these parameters we obtained the domain dictionaries of approximately 70 to 120 keywords. Such number of keywords is adequate for manual control and visual analysis used in our software.

Experiments with different sets of abstracts To evaluate the sensitivity of the combined measure to the broadness of the domain we used a document collection consisting of the abstracts and papers from two international conferences.

The first one, CICLing-2002 (Conference on Computational Linguistics and Intelligent Text Processing; `www.CICLing.org`) is a narrow domain-oriented conferencečiteGelbukh 2002. The document collection consisted of 48 abstracts (40 KB of text). The large and small domain dictionaries contained 115 and 74 keywords, respectively.

The second one, IFCS-2000 (International Federation of Classification Societies; `www.Classification-Society.org`) is a broad domain-oriented conference [6]. The document collection consisted of more than 200 abstracts. We eliminated from the collection all papers by invited speakers and the papers of invited sessions. The rest of the collection contained 166 abstracts (215 KB of

Table 2. Tuning the combined measure for clustering different data sets

α, β, p	$1, 0, -$	1,1,1	1,1,2	$1, \frac{1}{2}$,1	$1, \frac{1}{2}$,2	$\frac{1}{2}$,1,1	$\frac{1}{2}$,1,2	0,1,1	0,1,2
Abstracts of CICLing-2002, the nearest neighbor method									
LD	55%	**57%**	53%	**57%**	**57%**	55%	55%	36%	34%
SD	45%	**49%**	45%	47%	36%	36%	36%	34%	34%
Abstracts of CICLing-2002, the K-means method									
LD	46%	**48%**	46%	46%	**48%**	**48%**	**48%**	29%	33%
SD	52%	52%	52%	56%	**58%**	48%	54%	31%	42%
Abstracts of IFCS-2000, the nearest neighbor method									
LD	**47%**	**47%**	43%	38%	41%	34%	38%	40%	42%
SD	27%	35%	**38%**	35%	35%	27%	20%	24%	24%
Papers and abstracts of CICLing-2002, the nearest neighbor method									
Papers	**61%**	**61%**	57%	59%	57%	57%	55%	42%	34%
Abstracts	55%	**57%**	53%	**57%**	**57%**	55%	55%	36%	34%

text). The large and small domain dictionaries contained 107 and 70 keywords, respectively.[1]

As to the number of abstracts, it should be emphasized that we compare the results of automatic and manual clustering. When the number of documents exceeds 100-150 the expert's estimations are very fuzzy and so the contents of clusters become unstable. This is the reason for a limited number of papers in our experiments.

Experiments with full-text papers and abstracts For our experiments we used all 48 abstracts and papers (with the abstracts removed) of CICLing-2002 conference. The total size of the abstracts was about 40 Kb. For clustering abstracts we used only one (the best) dictionary contained 115 keywords. The total size of the papers was 1 Mb.

The dictionary for clustering papers had 187 keywords. Unlike the dictionary for clustering abstracts, this dictionary was constructed in a more traditional way: first an expert manually selected preferable words from the word frequency list and then assigned them the appropriate weights.

3.2 Experimental Results

Estimation of clustering quality We defined the clustering quality as coincidence of automatically selected clusters and the clusters selected by experts. For this we use the well-known formula to measure the similarity between two cluster sets [4]:

$$\chi = \max \frac{1}{N} \sum_{i,j}^{k} |A_i \cap B_j| \,,$$

[1] We thank the organizers of these conferences for providing us the corresponding materials.

where A_i, B_j, $i, j = 1, \ldots, K$ are two sets of clusters to be compared, K is the number of clusters, and N is the number of documents. So the quality χ is defined as the ratio between the number of equal documents in the closest clusters and the total number of documents.

In Table 2 we give the results of clustering of abstracts vs. full papers, of a narrow vs. broad domain-oriented conference, using the nearest neighbor vs. K-means method. In the tables LD stands for the large dictionary and SD stands for the small dictionary.

4 Conclusions and Future Work

Conclusions of the Experiments We have suggested a technique for clustering short texts, which is useful for clustering abstracts of scientific papers. Our experiments with abstracts suggest the following conclusions about the quality of the keyword lists used for clustering:

- The criterion of keyword selection we used provides stable contents of clusters with the combined measure. Namely, the difference in the clusters is about 10% to 20% when the dictionary varies in size in 50%, for different clustering methods, different broadness of the domain and combined measures.
- The size of the domain dictionaries affects differently the different methods of clustering. In particular, unlike the method of the nearest neighbor, the K-means method gives better results on a smaller dictionary.

As to the application of the combined document similarity measure, our experiments with abstracts suggest the following conclusions:

- For a narrow domain, the combined measure with the optimal selection of the parameters is better than the cosine measure in 5% to 10% and better than the polynomial measure in 20% to 40%.
- For a wide domain, the combined measure with the optimal selection of the parameters is better than the cosine and polynomial measures in 30%.

Comparing the results of clustering abstracts and full text papers for the narrow domain (the most interesting case), we found that with special precautions we have described here, abstracts can be clustered with almost as good results as the full texts. One should take into account the following: (1) With the traditional techniques these abstracts can be clustered with the accuracy no more than 40%–45%; (2) The agreement of the expert opinions is about 75%–80%.

Proposal on Open Access to Full Text Document Images Though one can achieve almost as good results on clustering abstracts as on clustering full texts of papers, still the results on the full texts are slightly better and can be achieved easier. To simplify the job of the search engines, both in search and in clustering the search results, especially in the context of the Semantic Web effort, we propose that digital libraries and Internet repositories *provide open access to document*

images of the papers. A document image is a vector of word frequencies, which can be restricted to a small list of keywords extracted from the whole document collection. This does not violate the copyright because it is impossible to recover full text of the paper from such a document image.

Future Work In the future, we plan to investigate various ways of constructing the keyword lists and apply different clustering methods. In particular, we will consider clustering the keywords to construct a new keyword space. We will also apply a stability-based criterion for determining the optimal number of clusters.

We plan to apply our techniques to very large medical database of Czech Ministry of Healthcare, in cooperation with our Czech colleagues.

References

1. Alexandrov, M., A. Gelbukh, and P. Makagonov (2000): *On metrics for keyword-based document selection and classification*. In: CICLing-2000, Proceedings of the 1st Intern. Conf. on Intelligent Text Processing and Computational Linguistics, Mexico, pp. 373–389.
2. Baeza-Yates, R., Ribero-Neto, B. (1999): *Modern Information Retrieval*. Addison Wesley.
3. Gelbukh, A., (ed.) (2002): CICLing-2002, *Computational Linguistics and Intelligent Text Processing*. Lecture Notes in Computer Science, N 2276, Springer-Verlag.
4. Hartigan, J. (1975): *Clustering Algorithms*. Wiley.
5. Hynek, J., Jezek, K., Rohlikm O. (2000): *Short Document Categorization—Itemsets Method*. In: PKDD-2000, Springer, LNCS, N 1910, 6 pp.
6. Kiers, H. et al., (eds.) (2000): *IFCS-2000, Proceedings of 7-th Intern. Conf. on Data Analysis, Classification, and Related Methods*. Studies in classification, data analysis, and knowledge organization, Springer-Verlag.
7. Makagonov, P., Alexandrov, M., Sboychakov, K. (2000a): *Keyword-based technology for clustering short documents*. In: Selected Papers. Computing Research, CIC-IPN, Mexico, pp. 105–114.
8. Makagonov, P., M. Alexandrov, K. Sboychakov (2000b): *A toolkit for development of the domain-oriented dictionaries for structuring document flows*. In: Data Analysis, Classification, and Related Methods, Studies in classification, data analysis, and knowledge organization, Springer-Verlag, pp. 83–88.
9. Makagonov, P. and Alexandrov, M. (2002): *Constructing empirical formulas for testing word similarity by the inductive method of model self-organization*. In: Advances in Natural Language Processing, Springer, LNAI, N 2379, pp. 239–247.
10. Manning, D., C. and Schütze, H. (1999): *Foundations of statistical natural language processing*. MIT Press.
11. Solomon, G. (1977): *Data dependent methods of cluster analysis*. In: Classification and Clustering, Academic Press, pp. 129–147.
12. Strzalkowski, T. (Ed.) (1999): *Natural Language and Information Retrieval*. Kluwer Academic Publishers.
13. Žižka, J., Bourek, A. (2002): *Automated Selection of Interesting Medical Text Documents by the TEA Text Analyzer*. In: A. Gelbukh (Ed.) Computational Linguistics and Intelligent Text Processing, CICLing-2002, Lecture Notes in Computer Science, N 2276, Springer-Verlag, pp. 402-404.