Clustering Abstracts Instead of Full Texts

Pavel Makagonov¹, Mikhail Alexandrov², and Alexander Gelbukh²,³

¹ Mixteca University of Technology, Mexico, Email: mpp2003@inbox.ru
² Center for Computing Research, National Polytechnic Institute, 07738, DF, Mexico Email: dyner@cic.ipn.mx, gelbukh@gelbukh.com, WWW: http://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 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

* 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.
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 for filtering out stopwords: only those words \( W \) are included in the DD for which

1. \( F_{\text{Dom}}(W) \gg F_{\text{Com}}(W) \); namely, \( F_{\text{Dom}}(W)/F_{\text{Com}}(W) > k \), where \( F_{\text{Dom}}(W) \) and \( F_{\text{Com}}(W) \) 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.