ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

*(An ISO 3297: 2007 Certified Organization)*

**Vol. 3, Issue 4, April 2015**

# Simple and Efficient Way to Cluster Documents for Growing Database

Dikhtiarenko Oleksandr[1], Biloshchytskyi Andrii[2]

Post Graduate student, Dept. of computer science fundamentals, Kyiv National University of Construction and Architecture, Kyiv, Ukraine[1]

Professor, Doctor of Technical Sciences, Head of Dept. of I.T., Kyiv National University of Construction and Architecture, Kyiv, Ukraine[2]

**ABSTRACT:** In this article we described a new method of clustering text documents. A frequency table of words from the documents was used as a characteristic of each document. These tables were created using term frequency which were cleaned from words that do not characterize a specific document and are common to the entire set of documents or for most of it. For the identification of such words, we calculated the percentage of documents in which this word occurs (inverse document frequency). The objectives of this publication were to determine the possibility of using frequency dictionary documents as their semantic characteristics and determine clustering method using frequency tables.

**KEYWORDS**: documents clustering; frequency word tables; semantic characteristics; clustering algorithms

## I. INTRODUCTION

While working on the search engine for finding fuzzy duplicates in electronic documents, it is necessary to speed up the search process. One way to do this is to reduce the total number of documents to be processed. This is possible if we group all documents, using the subject of the document itself. Unfortunately, we can not use this for document grouping branch of science, as different methods and models from one branch of science can be used quite successfully in another. For instance, some work in medicine may contain information models and methods of treatment for diagnosis of certain diseases, but the approach and method of data processing will be closer to information technologies than to health science. So it is more accurate to search for duplicates between other works in the field of information technologies, which use similar models and methods of data processing.

However, this work is an example that solves the problem only in the field of medicine. Therefore, to determine the scope of subjects covered by the scientific work, you can not use the attributes of the work itself, you need to focus on its content. A distribution of documents into groups without having preassigned categories is called clustering. For documents clustering we need to determine a method of obtaining quantitative data from the documents and determine the distance between these data. The next step is to determine at what distance the documents will get into one cluster, and if any - in various. One feature of this problem is a way to fill the documents database: first we have one document, then two and so on. We can not know in advance which documents get to our base and which groups can be in these documents.

## II. RELATED WORK

There are many ways to cluster the data or text documents, lets consider some of them. Latent Semantic Analysis [1] - a patented process (US Patent 4,839, 853), but the patent has expired already. The method is based on the construction of the document-term matrix, and then using them to simplify the singular value decomposition. A key feature of the algorithm is the recognition of elements of the same class in the absence of some of the attributes of this class in the document.

Probabilistic Latent Semantic Analysis [2] was developed by Hofmann in 1999, it was an improved version of the previous algorithm, s the incoming data used document-term matrix and the number of categories that need to share documents.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 4, April 2015

Hierarchical Agglomerative Clustering [3] - another document clustering method, which allows you obtain greater accuracy than the previous two. The essence of the algorithm is to bring together similar documents into scopes, and then merge the scopes. To determine the similarity of documents various techniques can be used , but it is best to use the scalar product which does not consider the size of the document. An illustrative example of association is shown in Figure 1.

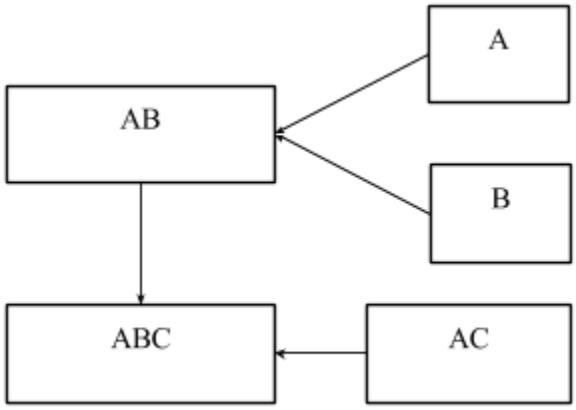


Fig. 1. Mergingofdocumentsbythegroups

K-means [4, 5] is a clustering method in which all documents are represented by points in a space. The centers of clusters were defined in the same space, and the algorithm attempts to minimize the distance from the center to each cluster member. The disadvantage of this method is the need to set the number of clusters as well as the definition of cluster centers. This method can not be used "on the fly" when we initially have only one document and then the number of documents in database is increasing.

Naive Bayes [6] is a very simple algorithm, which works by using probability. In this case, the document-term matrix is transformed into a probability (instead of the number of occurrences of the word it uses by the probability of the use of this word in this document or group). The disadvantage of this method it requires training data for all clusters, which is impossible in our case, since we do not even know how many clusters we will get.

N-grams [7] is not a clustering method and only affects the creation of document-term matrices. In the case of using this method, first it filters all stop words that do not carry the meaning out of context, and then it removes separating characters (as well as other non-letter symbols) and lastlyit then uses stemming for all words (cuts the endings). From the resulting string of characters, stands out all possible substrings of a given length (e.g. 3 characters), and it builds document-term matrix using the data it received. The resulting matrix is treated in another way, for example using k-means.

Self-Organizing Maps [8] is a method of projecting a multidimensional space onto a space with a small number of measurements, most often two or three-dimensional. For the projection of the multidimensional space, it uses a self-organizing map, which consists two parts: a neuron having a number of inputs equal to the dimension of the incoming data, and a coordinate in space, which projects data. The downside of this method is the result largely depends on the originally specified data (maps).

Orthogonalization (Polar Orthogonalization) [9] is the process of constructing for a given linearly independent system of vectors in a Euclidean or Hermitian space V is an orthogonal system of nonzero vectors that generate the same subspace of V. In theory and in some practical cases, it is much better than the LSA or k-means, but it requires a model for learning. Therefore, this method is not applicable to the particular case under consideration in this article.

All these methods are good and proven, but they have one drawback; they require initial data for training or initialization. The only exception is Hierarchical Agglomerative Clustering (HAC). In theory it is possible to create the

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 4, April 2015

initial cluster using HAC and then use the data to initialize the k-means or another method. However, in this work we will explore a different approach to the assessment of the contents of the document, and instead of document-term matrix will use frequency word tables.

## III. FREQUENCY WORD TABLES

In this paper the frequency word tables are tables consisting of two columns. The first column are the words that appear in the document after the pretreatment. Pre-treatment includes the elimination of various forms of expression, bringing all words in one form (canonization), eliminating synonyms and stop word, the stemming [10] can also be used. Stop words in this case are more than just a binder and pronouns. The table of word's frequencies do consider words that occur in all documents with great frequency. For example, the word "work" or "research" is likely to be encountered in scientific works, but by themselves (out of context) those words do not characterize the document from which they are taken. These words can not be determined if we use one or two documents. In this case, we also have to take a lot of documents before to compile a list of stop words.

Once the test document was processed, we created the table. After tables creation, the rows are sorted by the amount a word has been used, in descending order. From the resulting table we select a certain number of elements (n), which we assume to take as the document-specific characteristics, and later as the characteristic cluster. Table 1 is an example of such table.

Table 1. Exampleoffrequencywordtable

| № | Word itself | Frequency |
|---|---|---|
| 1 | word1 | 202 |
| 2 | word2 | 199 |
| 3 | word3 | 189 |
| 4 | word4 | 171 |
| … | | |
| n | wordn | 121 |

Apart from the n-th number of elements that are selected to characterize, we also take n elements starting from the last element (n + 1) and ending with the element number 2n. The second set is the "shadow" table, which is not to be used for work (the first comparison), but later both tables may exchange their items.

## IV. A WAY TO PROCESS FREQUENCY WORD TABLES

While working on the search engine for duplicates matching, we made some experiments [11] and made a conclusion that different contents of work (not containing duplicates) are written in one area can contain up to 80% of shared words. Thus, we had an idea that scientific papers which have been written in one area are likely to manipulate objects and terms that are specific to this area. In addition, these terms may be completely unnecessary in other areas. Therefore, comparing sets of words that occur in the document, we can determine how the subjects of these documents are close one to another. Let us consider an example, where we have an empty database for documents, and we get new documents one by one. For the first document we build a table T1 and as long as it is a single document, it immediately gets into a new cluster C1, and the characteristics of the document (frequency table and the shadow table) become a characteristic of the cluster itself. For each additional document we get, we build its comparison table (T2..Tn) and compare this table with all tables of all clusters we have(in this case only the table of the cluster C1). For comparison the tables, we used Jaccard distance [12]: $J(A,B) = \frac{|A \cap B|}{|A \cup B|}$.

The resulting value is compared to the minimum allowable value Jmin and if it is greater than this value, the document falls into the cluster C1, if not - it creates a new cluster C2, and the document is added to it. Each new document that enters the database is compared with all existing clusters and if none match it creates its own cluster. In

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 4, April 2015

order to adapt the characteristics of the cluster to all new documents which it contains, the frequency word table of cluster is recalculated each time the cluster gets a new document. For the recalculation, we use the new document frequency word tables of the new documents, and the number of documents in the cluster (k). Table T1 of the cluster contains the values { w1, w2, w3 … wn} where w - word with the assigned frequency, has a similar structure and table as the new document. When calculating the frequency of words we used the following formula: $w_{new} = w_c + \frac{w_{new}}{k}$, where $w_c$ is the word frequency from table of the cluster, $w_{new}$ is the word frequency from table of new document and k the number of documents that cluster contains already. At this stage we not only summarize the frequency of the main tables, but shadow tables as well. If the value of the frequency of some words from the shadow table is greater than the minimum value of the frequency in the main table, then the value from the shadow in the main table, pushes the lower value of the primary table into the shadow table. Thus, the frequency table always corresponds to the cluster set of documents that are in the cluster.

## V. EXPERIMENTAL VERIFICATION OF THE ALGORITHM

To check the capabilities of the algorithm we took a small database of scientific theses, which contained 150 documents. The small set of data was taken because we were manually verifying algorithm accuracy, and this is very hard to check for a large number of documents. The documents in the database had different topics (randomly selected from the total database of dissertations). The task for the algorithm was to accurately group the documents on the subject. In the experiment we used the following variables: N is the number of words in a frequency table (2N is the total number of words in the main and shadow tables), k is the minimum tolerance factor (Jaccard distance) in which two tables are similar.

To start, all documents were converted to text format, and punctuation and common stop words (work in identifying specific stop words were not conducted) were removed.

To assess quality, we used two parameters: precision (P) and the number of clusters (Q). The accuracy was determined for each cluster manually and estimated the number of documents of one subject to the total number of documents in the cluster, and then took the average value. The obtained results are listed in table 2.

Table 2. Testresults

| № | N | k | Q | P |
|---|---|---|---|---|
| 1 | 1000 | 0,1 | 5 | 7% |
| 2 | 1000 | 0,2 | 20 | 12% |
| 3 | 1000 | 0,3 | 96 | 37% |
| 4 | 500 | 0,1 | 7 | 11% |
| 5 | 500 | 0,2 | 42 | 76% |
| 6 | 500 | 0,3 | 102 | 91% |
| 7 | 200 | 0,1 | 6 | 7% |
| 8 | 200 | 0,2 | 40 | 43% |
| 9 | 200 | 0,3 | 107 | 88% |

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

*(An ISO 3297: 2007 Certified Organization)*

**Vol. 3, Issue 4, April 2015**

## VI. CONCLUSION

In this paper, we considered the method of clustering documents using frequency word tables. Typically document-term matrixes are used for clustering text documents, but it is impractical if we do not have a complete database of documents, and if the task is to analyze the new incoming documents. In this method, the frequency word table is constructed for each document only once, and the cluster tables are translated only if you enter a new document in the cluster, the others remain unchanged. The experimental data given to understand that the method can be used in practice, but the coefficients of the algorithm should be selected very carefully, which in turn requires further experimentation. Lastly, the data obtained should be cleaned from the stop words that can have a significant impact on the result of clustering.

## REFERENCES

1. Landauer, Thomas K., Peter W. Foltz, and Darrell Laham,"An introduction to latent semantic analysis"Discourse processes, Vol. 25, No.2-3, pp.259-284, 1998.
2. Heintze, Nevin, "Scalable document fingerprinting", USENIX workshop on electronic commerce, Vol. 3, No. 1, 1996.
3. Willett Peter, "Recent trends in hierarchic document clustering: a critical review", Information Processing & Management, Vol. 24, No.5,pp. 577-597, 1988.
4. Steinhaus H., "Sur la division des corps materiels en parties", Bull. Acad. Polon. Sci., Vol.4, pp.801—804, 1956.
5. Lloyd S., "Least square quantization in PCM's", Bell Telephone Laboratories Paper, Vol.28, No.2, pp.129-137, 1982.
6. Domingos, Pedro, Michael Pazzani,"On the optimality of the simple Bayesian classifier under zero-one loss", Machine learning, Vol. 29, No.2-3, pp.103-130, 1997.
7. Miao, Yingbo, VladoKešelj, EvangelosMilios, "Document clustering using character N-grams: a comparative evaluation with term-based and word-based clustering", In Proceedings of the 14th ACM international conference on Information and knowledge management, pp.357-358, 2005
8. Björck, Åke, Clazett Bowie, "An iterative algorithm for computing the best estimate of an orthogonal matrix", SIAM Journal on Numerical Analysis,Vol.8, No.2,pp.358-364, 1971.
9. Lovins, J. B., "Development of a stemming algorithm", MIT Information Processing Group, Vol.11, No.1-2, pp.22-31, 1968.
10. Biloshchytskyi A., Dikhtiarenko O.,"The effectiveness of methods for finding matches in texts", Management of complex systems, Vol.14, No.1, pp. 144 – 147, 2013.
11. Jaccard P.,"Distribution de la flore alpine dans le Bassin des Dranses et dansquelques regions voisines", Bull. Soc. Vaudoise sci. Natur.,Vol. 37, pp. 241–272, 1901.

## BIOGRAPHY

**Dikhtiarenko Oleksandr**is a Post Graduate student in the Department of computer science fundamentals ofKyiv National University of Construction and Architecture, Kyiv, Ukraine. He received Master of building technologies degree in 2012 from KNUBA, Kyiv, Ukraine. His research interests are approaches forfuzzy-search, data-mining.

**Biloshchytskyi Andrii**is a Doctor of Technical Sciences, Head of Department of information technologies ofKyiv National University of Construction and Architecture, Kyiv, Ukraine. He received his Ph.D. in project management from Kyiv National University of Construction and Architecture (KNUBA) in 2012. He has authored/co-authored more than 140 technical papers published in journals and a few books. His research interests are project management, information technologies, CAD systems.