ISSN 1392 – 124X INFORMATION TECHNOLOGY AND CONTROL, 2007, Vol. 36, No. 2

THE PECULIARITIES OF THE TEXT DOCUMENT REPRESENTATION, USING ONTOLOGY AND TAGGING-BASED CLUSTERING TECHNIQUE

Marijus Bernotas, Kazys Karklius, Remigijus Laurutis, Asta Slotkienė

Information Technology Department, Šiauliai University
Vilniaus St. 141, LT-76353 Šiauliai, Lithuania

Abstract. Text documents are very significant in the contemporary organizations, moreover their constant accumulation enlarges the scope of document storage. Standard text mining and information retrieval techniques of text document usually rely on word matching. An alternative way of information retrieval is clustering. In this paper we suggest to complement the traditional clustering method by document representation based on tagging, and to improve clustering results by using knowledge technology – ontology. The proposed method solves locally applied language incompact usage in the process of document clustering.

Keywords: document clustering, ontology, document management system.

1. Introduction

A search is usually applied for the information retrieval; this search retrieves documents from hierarchical systems according to the presence / absence of a word or a phrase independent of document features' matching. The user narrows a long list of results that satisfies his query after several search cycles. As well traditional search methods do not take into account the similarity of words and structure of documents within the corpus. The mentioned deficiencies can be overcome using document clustering method that, which has been very popular for a long time because it provides unique ways of digesting and generalizing large amounts of text documents. Without reference to the selected clustering method, the following steps are performed [9]: document representation selection, association measure selection, clustering method selection, cluster representation selection, validation of the results. Applying the usual clustering methods to text documents that are prepared in not prevalent languages as Lithuanian, Latvian, Estonian, Swedish, Dutch and similar, it was noticed that an uncompact dictionary of words and phrases is created, which is used for document representation. In order to solve this problem we applied tagging technology during the experiment, and for the improvement of clustering results we suggest to complement it by ontology.

2. Document Clustering

Document clustering can be determined as text clustering in a process of organizing pieces of textual information into groups which members are similar to some accordance, and groups as a whole are dissimilar to each other. Text clustering groups the documents in an unsupervised way and there is no label or class information. Clustering methods have to discover the relations between the documents and then, based on these relations, the documents are clustered. Given huge volumes of documents, a good document clustering method may organize those huge numbers of documents into meaningful groups, which enable further browsing and navigation of this corpus to be much easier.

Clustering of objects in the surroundings is an inborn feature of every human being. It is usual to categorize people to men and women intuitively, to distinguish things in accordance to geometrical shapes, destination, colors, etc. Document clustering has already been applied to information retrieval for thirty years. Research in the field has undergone a number of significant changes, from focusing on efficiency issues in the early years, to postulating the potential of clustering to increase the effectiveness of the information retrieval process [7] and it satisfies users' queries more.

Many clustering methods are described in the literature; these methods differ in document representation description, applied association measures, clustering algorithm and document groups' representation method. But given the high number and the strong diversity of the existing clustering methods, it is probably impossible to obtain a categorization that is both meaningful and complete. By focusing on some of the

discriminating criteria just mentioned we put forward the simplified taxonomy shown below [4]:

- Partitional clustering aims to directly obtain a single partition of the collection of items into clusters. Many of these methods are based on the iterative optimization of a criterion function reflecting the "agreement" between the data and the partition (Figure 1).
- Hierarchical clustering aims to obtain a hierarchy of clusters, called dendrogram that shows how the clusters are related to each other. These methods proceed either by iteratively merging small clusters into larger ones (agglomerative algorithms, by far the most common) or by splitting large clusters (divisive algorithms) (Figure 1).

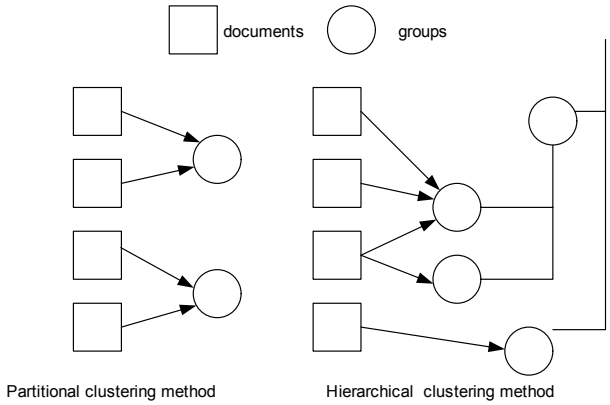

Figure 1. A simplified visual representation of different clustering algorithms

3. Document Representation Methods

The first step in the clustering process is to decide on the type and number of variables that describe each document. Documents are typically represented by a vector in an *n*-dimensional space, where *n* corresponds to the number of terms forming the indexing vocabulary of the database. In this model, the text documents are described by word vectors $<t_1, ...t_n>$, where w_i is the word's weight in the documents' collection, and n is the number of words in the dictionary of the document collection under clustering. The collection dictionary is created from words in the document text by performing the word filtering process; words of different form, semantic equivalents and such parts of speech as pronouns, prepositions and similar are rejected. It was noticed that the following characteristics are intrinsic to the vector space model (VSM):

1. Simple appliance for describing document features;
2. Easy to calculate similarity between two documents;
3. The clustering result is influenced by the collection dictionary that can ignore multi-words expressions, e.g. European Union. As well synonymous and polysemous interpretations are problematic as identical features can be assigned to them;
4. Nonoptimal appliance to local languages as Dutch, Latvian, Lithuanian and similar, as an uncompact dictionary is created and the clustering results can be perverted;
5. Word generalization relation is ignored for the retrieval of documents' features. Words in the collection dictionary are presented independently, however generally they belong to some group as, for example, gold and silver belongs to precious metals.

We suggest to solve the mentioned disadvantages (3-5) by applying a now popular technology of assigning tags to objects. This method is called tagging. Tags are metadata that describe contents of text documents; such metadata are also called keywords. The main advantage of the tagging method, which is useful in the clustering process, is a possibility to assign an object to several categories. The classification mechanism, based on object tagging, is applied to the description of various digital objects. It is usually performed by specialists of the subject area [5]. In the research performed by Golder [3] and other authors, it was noticed that if a possibility of assigning tags to objects is provided not to specialists, but the authors of these objects or other object users are provided with possibilities to assign their own tags and to validate the other, a very stable and regular classification system is created.

On the grounds of the tagging technology, the documents under clustering in the presented method will be described using tags assigned by the document authors. Sequentially, the documents are determined by the tags in the modified vector space model, where the document is equated to the multi-dimension vector with the vector tags $<t_1 \ldots. \ t_m >$, where w_j is a tag weight in the collection/collective document, while m is the number of tags present in the collection of the documents being clustered. As the tags are assigned by the document authors and their number can be different in the document collection, consequently, their weights w_j are normalized additionally in such way that the length of document vector would be equal to one. The formula

$$w_j = \frac{tf_j \times df_j}{\sqrt{\sum_{r=1}^{m} (tf_r \times df_r)^2}}$$

Is used, where tf_j (tf_r) is a frequency of a term (number of word occurrences in a document), while df_j (df_r) is a frequency of a document (number of documents containing the word).

Referring to the research results presented by Golder and Huberman, [3] we can affirm that a subjective and individual document tagging performed by its author has no influence on a good dictionary of the collection, as usually users select general terms that describe the document content.

4. Metrics comparison of the document clustering methods based on the different methods

During the experiment four text document collections of different size were created from the archive of 13769 papers (Table 1). Document collection was clusterized using different representation methods: by representing documents using the document text or using the tags assigned to it. The average number of tags assigned to a document varies from 1 to 7 tags. Five document classes have been compiled (culture, politics, sports, agriculture, law enforcement) in each of the document collections.

Table 1. The summary of the document collections characteristics used in the researches

Document collection	Quantity of the documents	Quantity of words in the document	Quantity of tags in the document
1	1056	36843	1379
2	2112	55352	1882
3	4224	81293	2449
4	8448	116755	2754

In order to estimate the similarity of documents in the clustering process the cosine coefficient is applied, the split K-means algorithm to five distribution is used. The obtained results are evaluated using two metrics:

- Entropy. It defines the degree of dispersion of documents of different classes in the limits of one distribution. The more the value of the entropy is closer to 0, the lesser is the dispersion degree, moreover, the better are the clustering results.
- Purity. It evaluates the degree of concentration of documents of one class in the limits of one distribution. The more the value of purity metrics is closer to 1, the better are the clustering results.

The entropy diagram (Figure 2) shows that the evaluation of the entropy is better if the document text is used instead of tags. Though as the number of documents increases the values of the entropy become similar. A small number of tags assigned to a document could have influence on its lesser value. And the value of the purity (Figure 2) in case of a larger correlation is better, when the clustering method based on tagging is used.

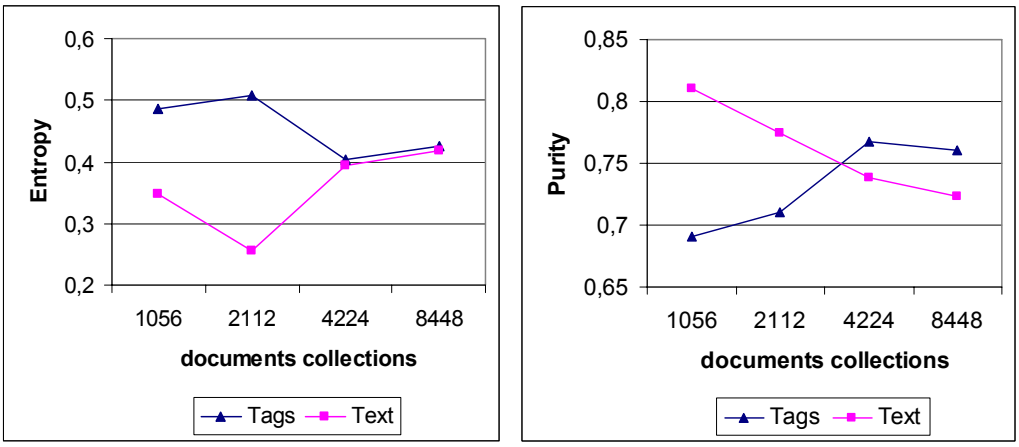

Figure 2. Documents collection clustering process of entropy and purity evaluation

Generalizing the results of the research, we can affirm that if an uncompact dictionary (words from the document text) based on the clustering method is used, quality results of clustering are obtained, however, if the number of documents in their collection is large, the tagging method is better in the purity metrics results.

5. Ontology Usage in the Clustering Method

Hotho and other authors [1,2,8] criticize the vector space model emphasizing that its representation used for the clustering methods is often unsatisfactory as it ignores relationships between important terms that do not co-occur literally. They suggest integrating the ontology into the vector space model. For information retrieval from text documents the ontology can be used as a conceptual specification of a document collection dictionary, which defines not only words or phrases, but also their semantic relationships: semantic similarity, interdependence definition, generalization relation.

Applying the ontology, the document would be described by a multidimensional vector that would encompass not only words in the document (tf defines the frequency of the word t_i in the document d), but also concepts (cf is the frequency of the concept c_j in the document):

$$\mathbf{t_d} = < tf(d,t_1)...tf(d,t_m), cf(d,c_1)...cf(d,c_l) >$$

Hotho and other authors [1,2,8] researches allow stating that knowledge appliance in the document representation improves the purity values. It is very significant to use the proposed method in the clustering method based on tagging as the document authors assign a limited number of tags (usually it does not exceed 10).

The principal idea of ontology-based clustering approach is the following:

- Use a simple ontology for generating alternative representations of the given document set;
- Representations are constructed by aggregating the original word vector into a concept vector, including information from the concept taxonomy;
- Standard k-means clustering is applied to the documents described by the reduced concept vector;
- Present clustering results using ontology net structure that varies in accordance to the content of the replaced tags.

The integration of ontology into the clustering method based on tagging provides a high level structural view for navigation through mostly unknown terrain and represents unstructured data (text documents) according to ontology repository.

6. Conclusions

Large document corpus may afford a lot of useful information to people. But it is also a challenge to find out the useful information from a huge number of documents. Information retrieval from the documents that are stored hierarchically in their storages is nonoptimal, as the document belongs only to one category. The archives of papers have a lot of documents that belong to more than one category due to their content. The appliance of one method of information retrieval – clustering – the mentioned disadvantage is eliminated. In our experiments we applied two of the clustering methods to the document collections, which differed in the way of document representation: words from the document text, tags assigned to the document by its author. The research results allow stating that the clustering method based on tagging effects on its results negatively if the collection of documents is of small scale, and provides quality results if the number of documents in the collection is large. The main reason is a small number of tags assigned to the document representation. On the grounds of the researches performed during the past years, we can affirm that this disadvantage can be overcome considering not only features that directly unify two documents, but also their context. The appliance of ontology enables this in the document representation stage, as the relationship between tags in the limits of the document distributions and document collections is being gathered, and a possibility to use the concept generalization relation is being acquired. We think that its integration would enable improving document similarities assessments as well.

Acknowledgements

This research was carried out during the project Software platform for remote workstations in editorial offices of the Eureka program with the co-financing of Agency for International Science and Technology Development Programmes in Lithuania

References

[1] **A. Hotho, A. Maedche, S. Staab, V. Zacharias.** On Knowledgeable Supervised Text Mining. *Proceedings of Text Mining Workshop, Springer*, 2002.

[2] **A. Hotho, S. Staab, G. Stumme.** Text Clustering Based on Background Knowledge**.** *Technical Report* 425, *University of Karlsruhe, Institute AIFB*, 76128 *Karlsruhe, Germany, April* 2003

[3] **S. Golder, B. A. Huberman**. Usage Patterns of Collaborative Tagging Systems. *Journal of Information Science*, *Vol*.32, *No*.2, 2006, 198–208.

[4] **D. Weiss**. Descriptive Clustering as a Method for Exploring Text Collections. *PhD thesis. Poznan University of Technology, Poznań, Poland*, 2006.

[5] **H-C. Chang, C-C. Hsu**. Using topic keyword clusters for automatic document clustering. *The Second International Workshop on Knowledge Discovery and Ontologies at ECML* 2005, *Vol*.1, 419- 424.

[6] **M. Bernotas, R. Laurutis, A. Slotkienė**. Issues on Forming Metadata of Editorial System's Document Management. *Information Technology And Control, Kaunas, Technologija,* 2005, *Vol*.34, *No*.4, 371 - 376.

[7] **N. Jardine, C. J. Van Rijsbergen.** The Use of Hierarchic Clustering in Information Retrieval. *Information Storage and Retrieval*, *Vol*.7, *No*.5, 1971, 217–240.

[8] **S. Bloehdorn, P. Cimiano, A. Hotho, S. Staab.** An Ontology-based framework for text mining. *LDV Forum – GLDV Journal for computational linguistics and language technology*, 2005, *Vol*.20, *No*.1, 87-112,

[9] **S. Theodoridis, K. Koutroumbas**. Pattern Recognition, Second Edition. *San Diego, Academic Press*, 2003.

Received April 2007.