Query Expansion and Evolution of Topic in Information Retrieval Systems*

Jiří Dvorský, Jan Martinovič, and Václav Snášel

Department of Computer Science, VŠB-Technical University of Ostrava,
17. listopadu 15, Ostrava - Poruba, Czech Republic
{jiri.dvorsky,jan.martinovic,vaclav.snasel}@vsb.cz

Abstract. Approach based on clustering will be described in our paper. Basic version of our system was given in [5] allows us to expand query through special index. Hierarchical agglomerative clustering of the whole document collection generates the index. Retrieving of topic development is specific problem. Standard methods of IR does not allow us such kind of queries for appropriate solution of information problem. The goal of presented method is to find list of documents that are bearing on topic, represented by user-selected document, sorted with respect to historical development of the topic.

1 Introduction

There are plenty of large text collections in the world. In connection with expansion of Internet these collections get bigger and bigger. Amount of processed data reach in present time dimensions that statistical properties of texts in collection become evident. This fact leads to new approaches, which involve methods from statistics, linear algebra, neural networks, and other (see [1, 9]).

Another important feature of these collections is their dynamic character. Modern surveys of information retrieval (IR) supposes that text collections have static character. Prior knowledge of topics' distribution is other presumption of current IR methods. Omission of these presumptions is more adequate to today's demands [2].

Retrieving of topic development is specific problem. Let's imagine that we want to perform query about war in Iraq from open source text collection. Set of terms contained in documents describing initial part of the war will be different from set of terms in document that characterize current state of war. Standard methods of IR [3] does not allow us such kind of queries for appropriate solution of information problem.

There are many IR systems based on Boolean, vector, and probabilistic models. All of them use their model to describe documents, queries, and algorithms to compute relevance between user's query and documents. Each model contains some constraints. Constrains cause disproportion between expected (relevant)

* This work was done under grant from the Grant Agency of Czech Republic, Prague No.: 201/03/1318

documents and documents returned by IR system. One of the possibilities how to solve the disproportion are systems for automatic query expansion, and topic development observing systems.

Approach based on clustering (see [7]) will be described in this paper. Basic version of our system was given in [5]. This version allows us to expand query through special index. Hierarchical agglomerative clustering of the whole document collection generates the index [6, 5].

Basic definitions of vector model will be briefly repeated in section 2. Section 3 contains introduction to cluster analysis, and description of agglomerative clustering. Section 4 is dedicated to query expansion algorithms that are based on work [5], including conclusions from tests. Description of topic development observing system, and its relationship to clustering algorithms are given in section 5. Section 6 gives us some conclusions and presents possibilities of future works.

2 Vector model

Vector model is dated back to 70th of the 20th century. The main goal of vector model is to enhance IR system based on Boolean model. Let's suppose vector IR system containing information about n *documents*. The documents are indexed by set of m *terms*. m dimensional vector represents each document in document collection, where every part of the vector corresponds to weight of particular term in given document. Formally:

$$d_i = (w_{i,1}, w_{i,2}, \ldots, w_{i,m}) \in \langle 0, 1 \rangle^m$$

Index of vector IR systems is the represents by matrix

$$D = \begin{pmatrix} w_{1,1} & w_{1,2} & \ldots & w_{1,m} \\ w_{2,1} & w_{2,2} & \ldots & w_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ w_{n,1} & w_{n,2} & \ldots & w_{n,m} \end{pmatrix} \in \langle 0, 1 \rangle^{n \times m}$$

Query in vector model is again m dimensional vector:

$$Q = (q_1, q_2, \ldots, q_m) \in \langle 0, 1 \rangle^m$$

Similarity between query Q and each document d_i can be computed as

$$Sim(Q, d_i) = \frac{\sum_{k=1}^{m} q_k w_{i,k}}{\sqrt{\sum_{k=1}^{m} (q_k)^2} \sqrt{\sum_{k=1}^{m} (w_{i,k})^2}}$$

There are many formulæ how to compute the similarity, we use one of the most frequent - cosine measure. The similarity can be understood as "distance" between query vector and vectors of documents in some vector subspace defined by matrix M.

Detail information about vector model can be found for example in [3].

3 Cluster analysis

The weight matrix M described above represents vast amount of numbers, that can be interpreted in some way. Among others, there is possibility to put documents together, which have approximately same coefficient of similarity to the potential query.

Finding of group of objects with the same or similar features within given set of objects i the goal of cluster analysis. These groups are called *clusters*. In other words, the group of similar objects forms the cluster.

Hierarchical clustering methods are important tool of cluster analysis. Hierarchical methods create hierarchy of clusters, grouped in cluster levels. The cluster levels arise during the computation and represents structure of hierarchy.

Hierarchical clustering methods can be divided into two groups:

agglomerative - At the beginning each object is considered as one cluster. Clusters are joined step by step together. The algorithm is over, when all objects form only one cluster.

divisive - The method works in reverse manner. At the beginning there is one cluster containing all objects. The clusters are sequentially divided until each cluster contains only one object.

Agglomerative clustering algorithm

1. **Create matrix of objects' distances**
   Matrix of objects' distances will be equal to term-document weight matrix D.
2. **Define each object as cluster**
   At the beginning each object is considered as one cluster i.e. there are as many clusters as objects. Sequentially, clusters are joined together and number of clusters drops down, when finally there is one cluster.
3. **Join pair of clusters with the least mutual distance**
   There many strategies, how to compute the distance. Among the most frequently used strategies belong:
   - *strategy of forthcoming neighbour* - Distances among all objects in two clusters are computed. The distance of clusters is then defined as minimal distance between any objects in these two clusters.
   - *strategy of farthermost neighbour* - Same as strategy of forthcoming neighbour, but maximum distance is chosen.
   - *average distance strategy* - Distances among all objects in two clusters are computed. The distance of clusters is then defined as average distance among any objects in these two clusters.
   - *median strategy* - The distance of clusters is then defined as median of distances among objects in these two clusters.
   - *Ward's method* - two cluster are joined together, if increase of sum of distances from centroid of clusters is minimal.

4. **recalculation of objects distance matrix**
   There are several strategies how to recalculate distance between new cluster, and other clusters:
   - Recalculation of all possible distances among new cluster and other clusters.
   - Already known distances between clusters, forming new cluster, can be exploited. Let c_3 be a cluster that is union of clusters c_1 a c_2. The distance of new cluster c_3 in respect of other clusters c_i can be determined as:

     $$d(c_3, c_i) = min(d(c_1, c_i), c(c_2, c_i))$$

     This recalculation strategy is faster than previous one, because there is no need of calculation of all distances for new cluster again.
5. **if there are more than one cluster, go to step 3**

---

4 Query expansion

Query expansion algorithms at first evaluate given query on collection of documents, and then select from relevant documents appropriate terms[1]. The original query is expanded with such selected terms. The expanded query is used to retrieve new set of relevant documents. The method is called feedback.

The feedback method has one important drawback. User query must be performed at first. And after searching, query is expanded with terms selected from retrieved documents. In our approach the relevant documents are replaced with the documents from cluster that is the most similar to the user query.

Two algorithms for query expansion were proposed (for details see [5]).

4.1 Description UP-DOWN-1 method

1. Algorithm begins at root of cluster hierarchy (cluster tree).
2. Similarity coefficient between the user query and the current cluster is calculated. If the similarity is greater than given threshold value, stop the algorithm and return current cluster.
3. If current cluster is a leaf in cluster hierarchy (i.e. the cluster contains only one object - document), stop the algorithm and return current cluster.
4. The similarity coefficients are calculated for both clusters that form current cluster.
5. The number of documents is determined in both clusters that form current cluster.

[1] What is appropriate term is another question. Good selection algorithm should prefer terms, that are specific for relevant documents to general terms in collection of documents. The selection algorithms therefore evaluate all terms in documents, considered in query expansion, and then they select terms with the highest value. For evaluation are often used Rocchio's weights, Robertson Selection Value, or Kullback-Leibler distance [4].

6. If the number is less than documents' number threshold, stops the algorithm, and return cluster with greater similarity coefficient.
7. In other case current cluster become cluster with greater similarity coefficient, and goes to step 2.

4.2 Description UP-DOWN-2 method

1. Algorithm begins at root of cluster hierarchy (cluster tree).
2. Similarity coefficient between the user query and the current cluster is calculated. If the similarity is greater than given threshold value, stop the algorithm and return current cluster.
3. If current cluster is a leaf in cluster hierarchy (i.e. the cluster contains only one object - document), stop the algorithm and return current cluster.
4. The number of documents is determined in both clusters that form current cluster.
5. *The algorithm goes to step 2 for all clusters with nonzero similarity coefficient, and with the number of documents greater than the threshold value.*

4.3 Experimental results

The UP-DOWN-1 method was tested at first. Original vector query, and vector query expanded with UP-DOWN-1 method was tested. Average improvement (rather impairment) of number of relevant documents can be seen at Fig 1.

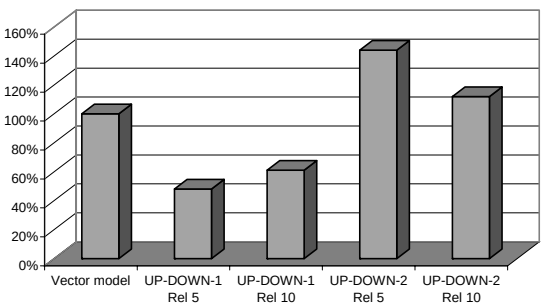

Fig. 1. Comparison of UP-DOWN-1 and UP-DOWN-2 methods

Graph 1 clearly shows that in case of using UP-DOWN-1 method, there is only 48% of relevant documents among the first five retrieved documents, with respect to the original vector query. Situation become better in case of the first ten documents, but only half of retrieved documents are relevant. The UP-DOWN-1 method does not ensure finding of the most similar cluster to the query, but it finds only one of similar clusters. After extensive exploration we

found out that the similarity coefficient partially depends on size of cluster. In other words document with less similarity to the query in one document cluster can have greater similarity coefficient than more similar documents in bigger cluster. In that case the query is expanded according to less similar cluster.

The UP-DOWN-2 method was proposed to eliminate this disadvantage. The method was tested on the same documents, and queries as UP-DOWN-1 method. The results can be seen at Fig 1. There can be seen that UP-DOWN-2 method gives much better results than UP-DOWN-1 method. There are three times more relevant documents in the first five documents with respect to the query expanded with the UP-DOWN-1 method. Moreover this method can find more relevant documents than original vector query.

5 Monitoring evolution of topic

Our research concern with he topics undergo an evolution. Let's assume document from collection of documents, that describes some topic. It is clear, that there will be some other documents in the collection that describes the same topic, but use different words to characterize the topic. The difference can be caused by many reasons. Among reasons belong evolution of topic in time. The first document about the topic use some set of words, that can change after some time period due to for example exploration of new circumstances, new fact, new political situation etc. Our experimental method should search an evolution for a given document and sort the result of the query.

Example 1. We want to search documents about an operation system. We have a document about the very latest operation system. Name of the operation system is changed during evolution. We want to search documents about all versions operation system.

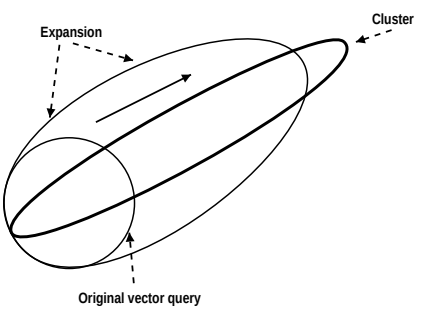

Fig. 2. Query extension

In the first way, vector query was expanded by terms from as close as possible cluster (see Figure 2). Experiments show, that our assumption was not correct.

Expanded vector query was not moved on clusters but only growing up (see Figure 3). This query expansion decreased coefficient of relevance, because non relevant documents are contained in result of expanded vector query.

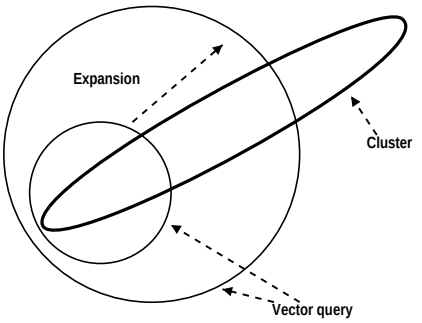

Fig. 3. Increase neighborhood

These experiments show that query expansion do not satisfy expectation. This results lead to drop this method whereas it leads to other method. The method finds cluster with similar documents and evolution of topic is examined in this cluster only. In the first instance we verify whether clusters include similar documents or not. And consequently when we obtain better result than vector query or not.

5.1 Experimental results

Collection of documents for testing purposes contain 1065 randomly selected documents from Parliament library from 1996 and 1998. The test consists of three steps:

1. Executing of vector query. Document 96-T0419 represents the query. Results of the query is given in table 1. The evolution of topic starts from document 96-T0419 i.e. query documents and ends in document 98-T0506 (see Figure 4).
2. Summarization[2] of documents in evolution of topic (see table 2).
3. Selection of the most important terms in documents from evolution (see table 3).

6 Conclusion

There are plenty of large text collections in the world. In connection with expansion of Internet these collections get bigger and bigger. Amount of processed data

[2] The summarization was done in MS Word. It is intended for checking the results.

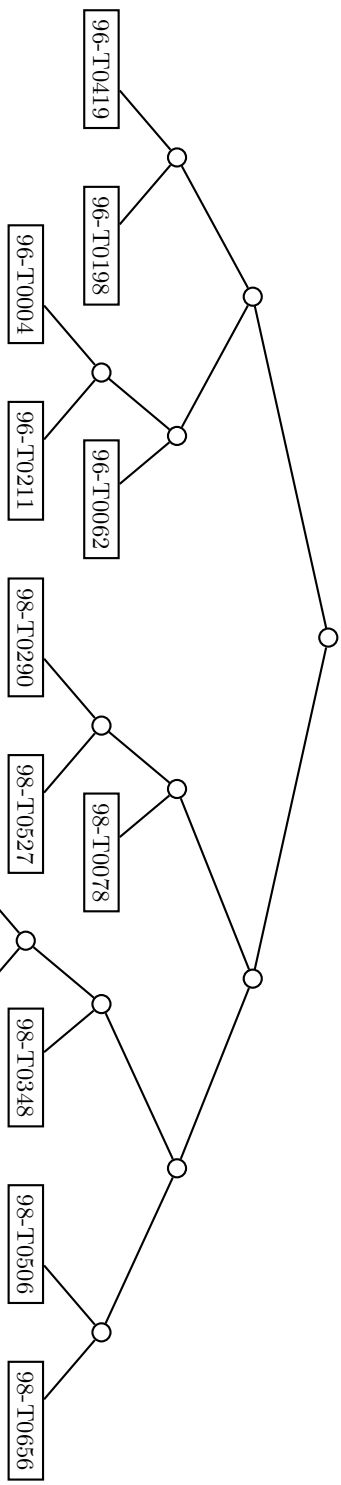

Fig. 4. Cluster tree corresponding to evolution of topic. Query document is the leftmost, and the last document in evolution is the rightmost.

Document	Similarity	Contained in topic evolution?
96-T0419	1.0000	yes
96-T0198	0.7550	yes
96-T0179	0.2385	
96-T0182	0.2766	
96-T0226	0.1514	
98-T0656	0.1290	yes
98-T0506	0.1098	yes

Table 1. Vector query results

96-T0419	na období 1998 - 2000PROTIDROGOVÉ POLITIKY VLÁDY NA OBDOBÍ 1998 - 2000Oblast snižování nabídky drog 93.3Oblast koordinace protidrogové politiky 213.5PROGRAM PROTIDROGOVÉ POLITIKY NA OBDOBÍ 1998 - 2000 305.1Oblast snižování nabídky drog 315.3- Počet předčasných úmrtí v důsledku užívání drog se na začátku 90. 3.2 OBLAST SNIŽOVÁNÍ NABÍDKY DROG
96-T0198	1. Rozsah užívání drog v ČR 2. Oblast snižování nabídky drog 1. Rozsah užívání drog v ČR Přibližně 10% dlouhodobých uživatelů drog je bez stálého bydliště. ČR je významnou tranzitní zemí kokainu. Opatření proti šíření a zneužívání drog v ČRNa tzv. 2. Oblast snižování nabídky drog drog
96-T0004	Zpráva o bezpečnostní situaci na území ČRMateriál se předkládá na základě usnesení vlády ČR č. 280Zpráva o bezpečnostní situaci na území ČR7. Bezpečnostní rizika na rok 1996 - shrnutí Zpráva o bezpečnostní situaci na území ČR/viz Příloha č. 1//viz Příloha č. 2
96-T0211	Zpráva o bezpečnostní situaci na území ČR v roce1996Materiál se předkládá na základě usnesení vlády ČR č. 308 2.5 Oběti trestné činnosti 5.2 Prevence kriminality Na objasněné trestné činnosti páchané na železnici se podíleli 40,23% (-13,08%).Na území hl. Počet trestných činů policistů vzrostl na 37,4 (+16,5%, +53 tr. č.).
96-T0062	května 1996 metodiku programů sociální prevence a prevence kriminality na místní úrovni, - přehled systému sociální prevence a prevence kriminality na ústředních orgánech státní správy.3. Plnění programu sociální prevence a prevence kriminality v předcházejícím obdobíPříloha č. 1 Schéma prevence kriminality f) projekty prevence kriminality na místní úrovni

Table 2. Summarization of documents in evolution of topic

96-T0419	drog drogách drogami drogové drogy oblast politiky prevence protidrogové snižování
96-T0198	drog drogách drogami drogové heroin pervitin prevence protidrogové především resocializace
96-T0004	bezpečnostní cizinců činů kriminalita kriminality migrace počet rizika trestné trestných
96-T0211	bezpečnostní činů kriminalita kriminality objasněnosti počtu policie trestné trestných zjištěných
96-T0062	kriminality prevence prevenci preventivních republikového sociální úrovni vnitra výboru východiska
98-T0290	malého malých podnikání podniků podporu podpory podpořeno projektů středního středních
98-T0527	čmzrb podnikatelského podnikatelský podporu podpory poskytnutí program programu projektu úvěru
98-T0078	doporučení malých měli míst podnicích podniků podniky pracovních pracovníků středních
98-T0218	bilance czechtrade deficitu dovozu egap exportu proexportní proexportních růstu vývozu
96-T0116	bilance deficitu dovozu firem obchodu především růst růstu vývozu zahraničního
96-T0348	aktivity exportní exportu obchodu podporu politiky proexportní vláda vývozu zahraničního
98-T0506	acquis evropské harmonizace legislativy oblast phare politiky programu přípravy vstupu
98-T0656	evropské nato politika politiky programovým průmyslu schválila systému vláda vládě

Table 3. The most important terms in documents contained in evolution of topic

reach in present time dimensions that statistical properties of texts in collection become evident. This fact leads to new approaches, which involve methods from statistics, linear algebra, neural networks, and other (see [1, 9]).

We can characterize a distribution of topics in a document with using clusters. These clusters are possible to use for the vector model and analysis how the topics undergo an evolution.

We developed the UP-DOWN-2 method which increase amount of founded relevant document.

In the future work we want to use large collection as a WebTrec for check whether this method. Next step we want to use Latent Semantic Indexing for computing document similarity [8].

References

1. Berry, M. W.; Browne, M.: Understanding Search Engines: Mathematical Modeling and Text Retrieval. SIAM Book Series: Software, Environments, and Tools, 1999.
2. Berry, M. W. (Ed.): Survey of Text Mining: Clustering Classification, and Retrieval. Springer Verlag 2003.
3. Baeza-Yates, R., Ribeiro-Neto, B.: Modern Information Retrieval. Addison Wesley, New York, 1999.
4. Carpineto C., de Mori R., Romano G., Bigi B.: An information-theoretic approach to automatic query expansion., in ACM Trans. Inf. Syst. 19(1), pp. 1-27, 2001
5. Dvorský J., Martinovič J., Pokorný J., Snášel V.: A Search topics in Collection of Documents. (in Czech), Znalosti 2004, in print.
6. Downs G.: Clustering in chemistry (an overview w.r.t. chemoinformatics), Math-FIT Workshop, Belfast, 27th April 2001 - Barnard Chemical Information Ltd.
7. Jain A., Dubes R.: Algorithms for Clustering Data. Prentice-Hall, Englewood Cliffs, NJ, 1988.
8. Praks, P., Snášel, V., Dvorský, J.: Latent Semantic Indexing for Image Retrieval Systems, SIAM LA, International Linear Algebra Society (ILAS), 2003.
9. Shumsky S., Yarovoy A.: Assocative searching of textual information, Neuroinformatics 1-99. MIFI, Moscow 1999.