APPROXIMATE REPRESENTATION OF TEXTUAL DOCUMENTS IN THE CONCEPT SPACE

Jasminka Dobša
University of Zagreb,
Faculty of Organization and Informatics
Pavlinska 2, 42000 Varaždin, Croatia
jasminka.dobsa@foi.hr

Bojana Dalbelo Bašić
University of Zagreb,
Faculty of Electrical Engineering and Computing
Unska 3, 10 000 Zagreb, Croatia
Bojana.Dalbelo@fer.hr

ABSTRACT

In this paper we deal with the problem of addition of new documents in collection when documents are represented in lower dimensional space by concept indexing. Concept indexing is a method of feature construction that is relying on concept decomposition of term-document matrix. This problem is especially interesting for application on World Wide Web. Proposed methods are tested for the task of information retrieval.
Vectors on which the projection is done in the process of dimension reduction are constructed on the basis of representations of all documents in the collection, and computation of the new representations in the space of reduced dimension demands recomputation of concept decomposition. The solution to this problem is the development of methods which will give approximate representation of newly added documents in the space of reduced dimension.

1 INTRODUCTION

In this paper we deal with the problem of addition of new documents in collection when documents are represented in lower dimensional space by concept indexing. This problem is especially interesting for application on World Wide Web. Proposed methods are tested for the task of information retrieval [1].
Methods for dimension reduction in the vector space model based on extraction of new parameters for representation of documents (feature construction) tend to overcome the problem of synonyms and polysemies which are two major obstacles in information retrieval. Our investigation is based on the method of feature construction called *concept indexing* which was introduced 2001 by Dhillon and Modha [6]. This method is relying on the *concept decomposition* of the term-document matrix.
Representation of new document in the vector space model is trivial. The problem appears when we want to add new documents in the space of reduced dimension. Namely, vectors on which the projection is done in the process of dimension reduction are constructed on the base of representations of all documents in the collection, and computation of the new representations in the space of reduced dimension demands recomputation of the concept decomposition. The solution to this problem is the development of methods which will give approximate representation of newly added documents in the space of reduced dimension.
Methods for addition of representations of new documents in the space of reduced dimension are already developed for LSI method [3,8]. The method of LSI was introduced in 1990 [4] and improved in 1995 [3]. Since then LSI is a benchmark in the field of dimension reduction. Kolda and O'Leary [7] developed a method for addition of representations of new documents for LSI method that uses semi-discrete decomposition.

2 DIMENSIONALITY REDUCTION BY THE CONCEPT DECOMPOSITION

Let the $m \times n$ matrix $A = [a_{ij}]$ be the term-document matrix. Then a_{ij} is the weight of the i-th term in the j-th document. A query has the same form as a document; it is a vector whose i-th component is the weight of the i-th term in the query. A common measure of similarity between the query and the document is the cosine of the angle between them. In order to rank documents according to their relevance to the query, we compute $s = q^T A$, where q is the vector of the query, while the j-th entry in s represents the score in relevance to the j-th document.

Techniques of feature construction enable mapping documents representations, which are similar in their content, or contain many index terms in common, to the new representations in the space of reduced dimension, which are closer than their representations in original vector space. That enables retrieving of documents which are relevant for the query, but do not contain index terms contained in the vector representation of query.

In this section we will describe the algorithm for computation of concept decomposition by the fuzzy *k*-means algorithm [5].

2.1 Fuzzy k-means algorithm

The fuzzy *k*-means algorithm (FKM) [9] generalizes the hard *k*-means algorithm. The goal of the *k*-means algorithm is to cluster *n* objects (here documents) in *k* clusters and find *k* mean vectors or centroids for clusters. Here we will call these mean vectors *concept vectors*, because that is what they present. As opposed to the hard *k*-means algorithm, which allows a document to belong only to one cluster, FKM allows a document to partially belong to multiple clusters. FKM seeks a minimum of a heuristic global cost function $J_{fuzz} = \sum_{i=1}^{k}\sum_{j=1}^{n} \mu_{ij}^{b} \left\| a_j - c_i \right\|$ where $a_j, j = 1,\ldots,n$ are vectors of documents, $c_i, i = 1,\ldots,k$ are concept vectors, μ_{ij} is the fuzzy membership degree of document a_j in the cluster whose concept is c_i and b is a weight exponent of the fuzzy membership.

2.2 Concept decomposition

Our target is to approximate each document vector by a linear combination of concept vectors. The *concept matrix* is an $m \times k$ matrix whose *j*-th column is the concept vector c_j, that is $C_k = [c_1, c_2, \ldots c_k]$. If we assume linear independence of the concept vectors, then it follows that the concept matrix has rank *k*. Now we define the *concept decomposition* P_k of the term-document matrix A as the least-squares approximation of A on the column space of the concept matrix C_k. Concept decomposition is an $m \times n$ matrix $P_k = C_k Z^*$ where Z^* is the solution of the least-squares problem, that is

$$Z^* = \left(C_k^T C_k\right)^{-1} C_k^T A . \qquad (1)$$

Z^* is a matrix of the type $k \times n$ and its columns are representations of documents in the concept space. Similarly, representation of query q in the reduced dimension space is given by $\left(C_k^T C_k\right)^{-1} C_k^T q$ and similarity between document and the query is given by the cosine of the angle between them.

Concept indexing is a technique of indexing text documents by using concept decomposition.

3 ADDITION OF NEW DOCUMENT REPRESENTATIONS IN THE CONCEPT SPACE

In this section novel algorithms for addition text documents representations in the concept space are proposed. The goal is to add new documents in a collection represented in the reduced dimension space, and this goal is achieved with and without an extension of the list of the index terms.

Let us introduce matrix notation that will be used in the section. Matrix

$$A = \begin{bmatrix} A_1 & A_2 \\ A_3 & A_4 \end{bmatrix} \qquad (2)$$

will be an extended term-document matrix, where A_1 a is matrix of starting documents in the space of starting terms, A_3 is a matrix of starting documents in the space of added terms, A_2 is a matrix of added documents in the space of starting terms and A_4 is a matrix of added documents in the space of added terms. Further, let m_1 be number of starting terms, m_2 number of added terms, n_1 number of starting documents and n_2 number of added documents.

Here we will introduce two methods of approximate addition of new documents in the concept space:

(a) projection of new documents on existing concept vectors (Method CI_A),

(b) projection of new documents on existing concept vectors extended in dimensions of newly added terms (Method CI_B).

Assume that documents of a starting matrix A_1 are clustered by fuzzy *k*-means algorithm and centroids of clusters are computed. Let C_1 be the concept matrix the columns of which are concept vectors and let C_2 be a matrix consisting of extensions of concept vectors in dimensions of added terms. Extensions of concept vectors are calculated analogously as concept vectors by using respective columns of matrix A_3 instead of columns of A_1. Let extensions of concept vectors form extension of the concept matrix denoted by C_2. Then $C = \begin{bmatrix} C_1 \\ C_2 \end{bmatrix}$ is the concept matrix the columns of which are concept vectors extended in dimensions of newly added terms. Representations of documents in the concept space of extended term-document matrix will be given by expression

$$\left(\begin{bmatrix} C_1 \\ C_2 \end{bmatrix}^T \begin{bmatrix} C_1 \\ C_2 \end{bmatrix}\right)^{-1} \begin{bmatrix} C_1 \\ C_2 \end{bmatrix}^T \begin{bmatrix} A_1 & A_2 \\ A_3 & A_4 \end{bmatrix}$$
$$\approx [\left(C_1^T C_1\right)^{-1} C_1^T A_1 + \left(C_1^T C_1\right)^{-1} C_2^T A_2 \qquad (3)$$
$$\vdots \left(C_1^T C_1\right)^{-1} C_1^T A_3 + \left(C_1^T C_1\right)^{-1} C_2^T A_4]$$
$$= [(4)+(5) \ \vdots \ (6)+(7)].$$

This expression can easily be shown if approximation $\left(C_1^T C_1 + C_{21}^T C_2\right) \approx C_1^T C_1$ is assumed. Such an approximation is justified by the fact that extensions of concept vectors are sparser than concept vectors formed from starting documents, because the coordinates of extended concept vectors are weights of added terms which were not included in list of the index terms before addition of new documents. It was established, by experiment, that $\left\|C_2^T C_2\right\|_2 << \left\|C_1^T C_1\right\|_2$. The number of operations is significantly reduced by this approximation, because inverse $\left(C_1^T C_1\right)^{-1}$ is already computed during the computation of starting documents projection.

This approximation is not necessary for the application of Method CI_A, because this method does not use extensions of concept vectors. Representations of starting documents are given by expression (4), while representations of added documents are given by expression (5) in the last row of expression (3).

By the Method CI_B added documents are projected on the space of extended concept vectors. Vector representations of starting documents are already known, and they are given by the (4) in the last row of expression (3), while representations of added documents are computed by the formula

$$\left(C_1^T C_1\right)^{-1} C_1^T A_3 + \alpha\left(C_1^T C_1\right)^{-1} C_2^T A_4, \quad (8)$$

where coefficient $\alpha > 1$ has a role of stressing the importance of added terms and documents.

4 EXPERIMENT

Experiments are conducted on MEDLINE collection of documents. The collection contains 1033 documents (abstracts of medical scientific papers) and 35 queries. The documents of collection are split randomly into two parts: starting documents and added documents. The ratio of starting and added documents is varied: first added documents form 10% of the whole collection, then 20% of the whole collection, and so on. Starting list of index terms is formed on the basis of starting collection of documents. In the list are included all words contained in at least two documents of starting collection, which are not on the list of stop words. Further, the list of index terms is formed for the whole collection of documents in an analogous way. The obtained list of index terms for the whole collection contains 5940 index terms.

We have used measure of mean average precision (MAP) for evaluation of the experimental results. Concept decomposition is conducted under starting collection of documents and added documents are represented in the concept space by using one of the described methods for approximate addition of documents. After that, an evaluation of information retrieval performance is conducted under the whole collection of documents. Dimension of the space of reduced dimension is fixed to *k=75*.

In the second row of Table 1, there is MAP of information retrieval in the case that procedure of concept decomposition is conducted under whole collection of documents (percentage of added documents is 0%). This value presents MAP in the case of recomputation of concept decomposition when new documents are added in the collection. All other values of MAP in the cases when the collection is divided into collection of starting and added documents in the different ratios, could be compared to this value.

The second column of Table 1 presents number of added documents, while the third column presents number of added terms. Let us note that number of added terms grows linearly, and that the collection with only 20% of starting documents is indexed with a much smaller set of index terms then the whole collection. The fourth row presents MAP of information retrieval for approximate addition of new documents by Method CI_A (without addition of new index terms). The rest columns of Table 1 present MAP of information retrieval for approximate addition of new documents by Method CI_B (with addition of new index terms) for different values of parameter α.

From the results we can conclude that an addition of new index terms does not improve results of MAP significantly. The results obtained by Method CI_B, α=2.0 are not significantly better in comparison to results obtained by Method CI_A according to pared t-test (α=0.05).

5 CONCLUSIONS

Values of MAP for approximate methods are acceptable in comparison to repeated computation on concept decomposition when the number of added documents is the same or smaller than the number of starting documents. There is a drop of MAP when the number of added documents exceeds the number of starting documents.

Results of MAP are not significantly improved by the methods that use extended list of index terms obtained as a result of addition of documents. It is interesting to notice that this statement is valid even in the cases when the list of index terms is significantly enlarged, which is when larger proportion of documents is added. This results show a great redundancy present in the textual documents.

Percentage of added documents	**Number of added documents**	**Number of added terms**	**MAP** Method CI_A	**MAP** Method CI_B α=1.0	**MAP** Method CI_B α=1.5	**MAP** Method CI_B α=2.0
0	0	0	54.99	54.99	54.99	54.99
10	104	456	51.98	52.20	52.33	**52.37**
20	208	753	54.96	55.10	55.09	**55.23**
30	311	1264	51.90	51.78	51.97	**52.03**
40	414	1673	50.84	50.60	51.09	**51.64**
50	517	2089	48.64	47.99	48.29	**48.64**
60	620	2696	44.26	44.08	45.04	**45.49**
70	723	3282	43.59	41.86	42.32	**42.70**
80	826	4024	39.87	40.56	42.56	**43.74**

Table 1. *MAP of information retrieval for methods of approximate representation of added documents by projection on existing concept vectors (Method CI_A) and extended concept vectors (Method CI_B). The best results for every split of the collection are achieved by using Method CI_B, α=2.0 (shown bolded).*

References

[1] R. Baeza-Yates, B.Ribeiro-Neto. *Modern Information Retrieval*, Addison-Wesley, ACM Press, New York, 1999.

[2] M. W. Berry, Z. Drmač, E. R. Jessup. Matrices, Vector Spaces, and Information Retrieval, *SIAM Review*, Vol. 41. No. 2. 1999, pp. 335-362.

[3] M. W. Berry, S. T. Dumais, G. W. O'Brien. Using linear algebra for intelligent information retrieval, *SIAM Rewiew,* Vol. 37. 1995, pp. 573-595.

[4] S. Deerwester, S. Dumas. G. Furnas. T. Landauer, R. Harsman. Indexing by latent semantic analysis, *J. American Society for Information Science,* Vol. 41. 1990, pp. 391-407.

[5] J. Dobša, B. Dalbelo-Bašić. Concept decomposition by fuzzy k-means algorithm, *Proceedings of the IEEE/WIC International Conference on Web Intelligence, WI 2003,* 2003, pp. 684-688.

[6] I. S. Dhillon, D. S. Modha, Concept Decomposition for Large Sparse Text Data using Clustering, *Machine Learning* , Vol. 42. No. 1, 2001, pp. 143-175.

[7] T. Kolda, D. O'Leary. A semi-discrete matrix decomposition for latent semantic indexing in information retrieval, *ACM Trans. Inform. Systems,* Vol. 16, 1998, pp. 322-346.

[8] G.W. O'Brien. *In formation Management Tools for Updating an SVD-Encoded Indexing Scheme,* Master s thesis, The University of Knoxville, Tennessee, 1994.

[9] J. Yen. R. Langari. *Fuzzy Logic: Intelligence, Control and Information,* Prantice Hall, New Jersey, 1999.