Text retrieval using self-organized document maps

Krista Lagus

Helsinki University of Technology
Neural Networks Research Centre
P.O. Box 5400, FIN-02015 HUT, Finland

Abstract

A map of text documents arranged using the Self-Organizing Map (SOM) algorithm (1) is organized in a meaningful manner so that items with similar content appear at nearby locations of the 2-dimensional map display, and (2) clusters the data, resulting in an approximate model of the data distribution in the high-dimensional document space. This report describes how a document map that is automatically organized for browsing and visualization can be successfully utilized also in speeding up document retrieval. Furthermore, experiments on the well-known CISI collection indicate improved performance compared to Salton's vector space model and to Latent Semantic Indexing, measured by average precision when retrieving a small, fixed number of best documents.

Keywords: information retrieval, SOM, text mining, document maps, LSI

1 Introduction

People approach large information spaces at least with the following different kinds of motives: (1) to search for a specific piece of information or for information of a specific topic, (2) to gain familiarity with or an overview of some general topic or domain, or (3) to locate something that might be of interest, without a clear prior notion of what "interesting" should look like. The field of Information Retrieval (IR) (Salton and McGill, 1983; Baeza-Yates and Ribeiro-Neto, 1999) develops methods that focus on the first situation, whereas the latter motives are mainly addressed in approaches that concentrate on exploration and on visualizing properties of data (for an overview, see Hearst, 1999). Often the motives and information needs of users in actual sessions alternate. Consequently, a balanced combination of visualization, exploration, and search tools and facilities is likely to be more useful in real systems than any single approach alone.

It has been shown previously in the WEBSOM project (Honkela et al., 1996; Kaski et al., 1998a; Lagus et al., 1999; Kohonen et al., 2000) that using the Self-Organizing Map (SOM) algorithm (Kohonen, 1982; Kohonen, 1995) very large text collections can be automatically organized onto *document maps* that are suitable for visualization and intuitive exploration of the information space. In a user study reported in (Chen et al., 1998) similar maps were found especially useful for a situation in which the users were mainly browsing for something interesting without a clear information need.

The construction and use of exploration models and search indices consumes processing time, memory, and disk space. Especially with very large collections that push the limits of computer systems it is an advantage if a single set of data structures or indices can be utilized for all of the identified tasks related to text mining. Furthermore, in real systems any search and exploration methods must be computationally efficient. In particular, the delay perceived by users is critical. It is therefore important to develop methods that can speed up the search process while maintaining high perceived quality, particularly in the range of high precision and low recall which is most crucial in actual user settings.

The two aspects of the *vocabulary problem* are well-known: a single term may have many different meanings, and the same idea may be expressed in various ways, e.g. by using different terms. This can be viewed as a problem of noisy data on one hand, and of missing data values on the other hand. In general, such problems can be attacked by utilizing an appropriate *model* of the data space, created with methods that reduce noise and cope gracefully with missing values and partial matches.

In this article a computationally efficient method is presented that is suitable for performing information retrieval using document maps that were organized for exploration of document collections. Furthermore, experiments on the well-known CISI reference collection show that comparable or improved retrieval performance is obtained when compared to two standard methods, Salton's vector space model (Salton et al., 1975) and the LSI (Deerwester et al., 1990).

The organization of the paper is as follows: The introduced method is presented in Sec. 2, the experimental setup in Sec. 3 and the results in Sec. 4. First, the standard algorithms used in the experiment are briefly described.

1.1 The Self-Organizing Map (SOM)

The SOM (Kohonen, 1982; Kohonen, 1995) is a nonlinear projection method that maps an originally high-dimensional data space onto a usually two-dimensional map grid in an orderly fashion. The map grid locations are associated with so-called reference vectors that act as *local models* of the closest data items and, to a lesser degree, of the neighboring map regions. By virtue of the SOM algorithm nearby map locations contain similar data items, enabling intuitive visualization of the data space. Furthermore, the reference vectors partition the data into subsets of similar items, performing a *clustering* of the data.

Numerous applications of the SOM are described in (Kohonen, 1995); for an extended list see (Kaski et al., 1998b).

1.2 Salton's vector space model (VSM)

In the vector space model (Salton et al., 1975) documents are represented as points in a t-dimensional space where t is the size of the vocabulary, and the component of the document vector d_{ij} reflects the frequency of occurrence of the term i in the document j. Furthermore, the terms can be weighted in various ways, e.g. using a $tf \times idf$ scheme (Salton and Buckley, 1987).

1.3 Latent Semantic Indexing (LSI)

Global relationships between words can be deduced from their co-occurrence patterns across documents. A method called Latent Semantic Indexing (LSI) (Deerwester et al., 1990) utilizes this property by applying singular-value decomposition to the document-by-word matrix to obtain a projection of both documents and words into a space called the *latent space*.

Dimensionality reduction is achieved by retaining only the latent variables (projection dimensions) with the largest variance (largest eigenvalues). Subsequent distance calculations between documents or terms are then performed in the reduced-dimensional latent space.

2 Document retrieval using the SOM

If the object of the search is stated as locating a small number n of best documents in the order of goodness corresponding to a query, the following search strategy can be applied:

1. Indexing phase: Apply the SOM to partition a document collection of D documents into M subsets or clusters. Represent each subset by its centroid.
2. Search phase: For a given query,
   (a) *Pre-select:* Based on comparison with the centroids, select the best subsets and collect the documents in those subsets until K documents ($K >= n$) are obtained.
   (b) *Refine:* perform an exhaustive search among the K prospective documents and return the n best ones in the order of goodness. The exhaustive search can be carried out with identical document encoding as used in forming the subsets.

The hierarchical strategy has two effects: the complexity is reduced compared to a full search, and the returned set is different from that of an exhaustive search.

One can view the role of the parameter K as determining the balance between the clustering and the refining search: with a small K a small number of clusters is trusted to suffice for pre-selecting the best documents. With a larger K a greater number of clusters will be selected to be screened closely by the detailed search. Furthermore, when SOM is used for obtaining the subsets, the clustering is implicit (i.e. a cluster may be formed by several neighboring map units) and the appropriate value of K is likely to depend on the stiffness of the map, which is affected by the SOM learning parameters.

Effects on computational complexity. The partitioning of the document collection can be constructed offline. In general, obtaining a meaningful partitioning of a large data set is computationally hard. However, by utilizing the SOM the task can be performed fast (Kohonen et al., 2000).

In carrying out a single search, an upper limit of $O(M + K + n \log n)$ comparisons of document vectors are needed (M for the comparison with centroids, K for comparison with the result set of documents, and $n \log n$ for retaining the best units in a heap data structure). In contrast, the complexity of an exhaustive search would be $O(D+n \log n)$, D for comparing the query with each document, and $n \log n$ operations for saving results in a best-first order.

Effects on search quality. If tight clusters of similar documents are obtained, the data is well characterized by the centroid of the subset, and thus the computation of similarities with respect to the centroids instead of to individual documents is likely to form a good approximation of full search. Furthermore, a clustering may bring related documents together based on *family resemblance*, i.e. the cluster may not be defined in terms of a single set of common traits. Instead, as in a family, all members resemble at least some other members of the family, but the features in which resemblance occurs may differ. As a consequence, through clusters the query can reach documents that share few or no words with the query.

A straightforward approach for utilizing clusters in searching would be to return for a query all documents in the best-matching clusters. However, if a topical clustering is used, this approach is likely to cause also false hits to appear high in the list (this was confirmed in a preliminary experiment not reported here). Instead, the document map can be thought of as a *domain model* that is utilized for determining the most central topical clusters for a query. Subsequently the best clusters are scanned more closely to identify the best hits for the particular query. The pre-selection stage is likely to discard outlier (one-of-a-kind) hits, as well as a large number of documents that only weakly resemble the query (e.g. due to terms with small weight in common with the query). The further detailed search is thus made easier, and hence retrieval results may even improve.

Since the document map organizes and clusters documents based on their overall similarity, the map can be viewed to perform a kind of indirect disambiguation for the words in the documents, and subsequently on the query as well.

3 Experiments

3.1 Data set

The CISI collection (CISI-collection, 1981) was selected for the experiments in order to evaluate the method on a standard test collection where the relevant documents are known and the properties of the collection are well-understood by practitioners of the field. The collection consists of 1460 documents and 76 information requests or *queries* with an average of 41 known relevant documents per query. The average lengths of documents and queries were 115 and 51 words, respectively.

3.2 Evaluation measure

The evaluation techniques used in IR are based on the notions of *precision* and *recall*, defined as follows:

$$\text{precision} = \frac{\text{number of relevant items retrieved}}{\text{total number of items retrieved}} \tag{1}$$

$$\text{recall} = \frac{\text{number of relevant items retrieved}}{\text{number of relevant items in collection}} \tag{2}$$

The retrieval performance was compared using *non-interpolated average precision over all relevant documents* (AP) (Voorhees and Harman, 2000). AP is calculated for a single query by recording the precision each time a relevant document is retrieved, and taking an average of these (the precision of non-retrieved documents is calculated as zero). The results are further averaged over all queries. The measure not only takes into account the *number* of relevant documents retrieved, but also their *order* in the result set, rewarding systems for ranking relevant documents high in the returned lists of results.

3.3 Searches using the document map.

In order to evaluate the retrieval performance of document maps constructed for exploration of a collection, the document map was created following the principles previously established in the WEBSOM project (for a state-of-the art description, see Kohonen et al., 2000).

Preprocessing and document encoding. The texts were preprocessed by replacing words with their base forms using the TWOL morphological analyzer (Koskenniemi, 1983), and by removing a stoplist of common, uninformative words. Terms occurring less than five times were discarded as well. The final vocabulary consisted of 2136 terms. The documents were encoded as vectors using the vector space model where a $tf \times idf$ weighting was applied.

Construction of the map. The 1460 document vectors were organized on a SOM of 150 units (10×15) according to principles established in the WEBSOM project. The map size was selected as appropriate for browsing, so that each map unit would contain an average of 10 documents. Furthermore, the so-called dot product SOM was utilized (Kohonen, 1995) in which distance computations are performed using dot products on normalized document vectors (cosine distance measure).

Performing searches on the map. For each query, the documents in the best map units were pre-selected until a list of K documents was obtained. To return the final n documents a refinement search was performed: out of the K pre-selected documents the n most similar were chosen based on the cosine measure.

Estimation of the value for K. To choose a value for K the test data (the set of 76 queries) was randomly split into two halves with equal probability. One of the halves was used to select a value for K while the other was set aside and used for comparison of the different methods. The different values of K that were tried and the respective performances are shown in Figure 1. The value $K = 300$ was chosen.

3.4 Searches using VSM and LSI

In experiments with the VSM and the LSI the preprocessing, the vocabulary of terms, and the weighting of words were identical to those of the document map experiment to avoid comparing different preprocessings. With both models the searches were carried out by comparing the query vector with each document vector (in the vector space or in the latent space), and by selecting the n best documents.

In the VSM experiment the distances were calculated as dot products on normalized document and query vectors.

The dimensionality of the latent space in LSI was estimated in the same manner as the parameter K in the document map experiment, i.e. based on results obtained with the first half of the test set. The dimensionalities 100, 120, 130, 140, 150, and 160 were tried, of which 140 performed best and was selected for comparison of methods.

4 Retrieval results

Each method was asked to return the n best documents, with various recordings of n. The results presented in Figure 2 indicate that the document map performs somewhat better than either the VSM or the LSI.

5 Conclusions and discussion

The study shows that a document map that was created for interactive exploration of a text collection can be successfully utilized as a clustering in speeding up document retrieval. Fur-

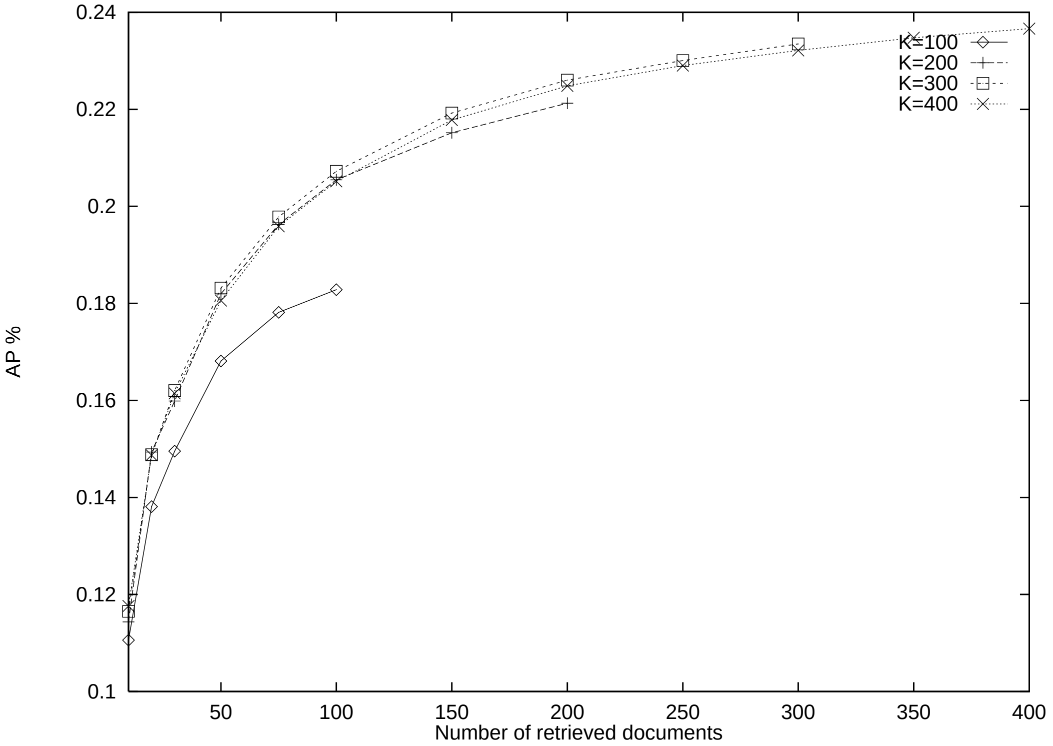

Figure 1: Selection of the value for parameter K. Average precision (AP) was calculated using the suggested document map method with various values of K (the number of pre-selected documents), and varying the number of retrieved documents (n). The results are averaged over the queries in the first half of the test set. The AP for larger n is higher as a larger number of relevant documents fall into the result set.

thermore, the experiments suggest that using document maps an equal or improved retrieval performance may be achieved compared to standard methods such as VSM or LSI, with reduced computational complexity.

Experiments with other well-understood collections are needed to confirm the results as well as to examine further the conditions in which the proposed method improves performance. Additionally, the approach should be studied with very large collections.

Still further improvement, perceived by a user but not readily measurable within the IR framework, may be achieved by visualizing the locations of the best hits on the map display and by enabling exploration of the map. In such an environment the user may visually confirm the results and interactively refine the search, as well as locate further interesting documents identified by resemblance with some of the retrieved ones.

References

Baeza-Yates, R. and Ribeiro-Neto, B., editors (1999). *Modern Information Retrieval.* Addison Wesley Longman.

Chen, H., Houston, A. L., Sewell, R. R., and Schatz, B. R. (1998). Internet browsing and searching: User evaluations of category map and concept space techniques. *Journal of the American Society for Information Science (JASIR)*, 49(7):582–603.

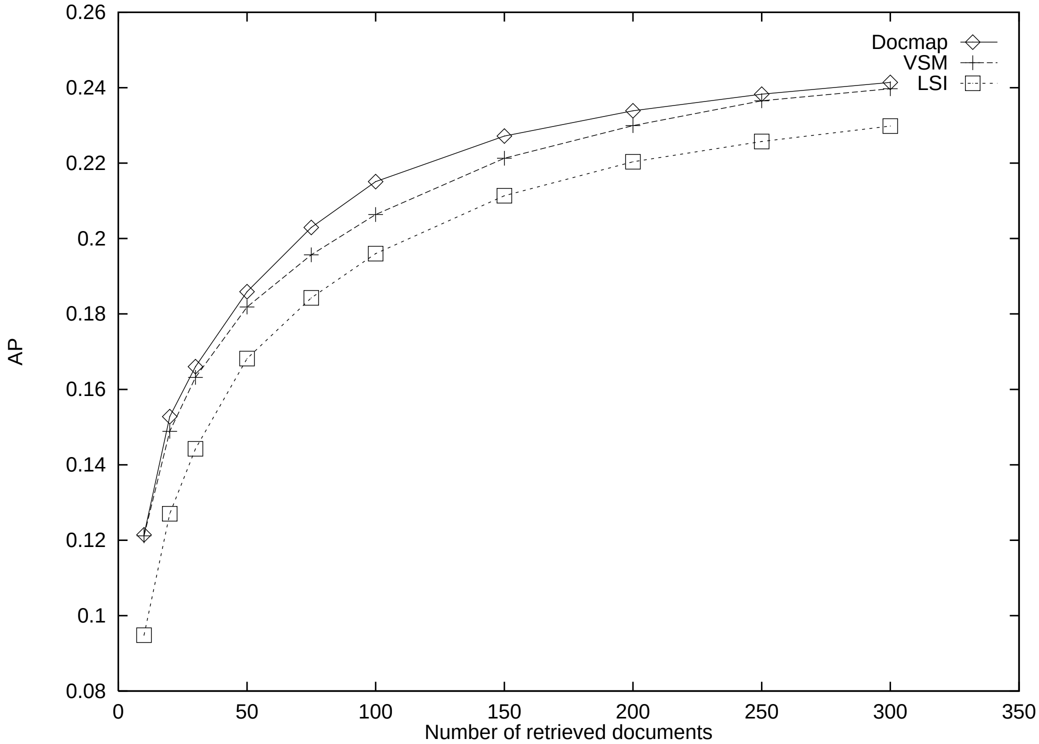

Figure 2: Comparison of methods. Average precision (AP) was calculated using the suggested document map method (Docmap), the Salton's vector space model (VSM) and the Latent Semantic Indexing (LSI) with various numbers of retrieved documents. The results are averages over the queries in the second half of the test set.

CISI-collection (1981). The CISI reference collection for information retrieval. 1460 documents and 76 queries. http://local.dcs.gla.ac.uk/idom/ir_resources/test_collections/cisi/.

Deerwester, S., Dumais, S. T., Furnas, G. W., and Landauer, T. K. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41:391–407.

Hearst, M. A. (1999). *Modern Information Retrieval*, chapter 10. User Interfaces and Visualization, pages 257–324. Addison Wesley Longman.

Honkela, T., Kaski, S., Lagus, K., and Kohonen, T. (1996). Newsgroup exploration with WEBSOM method and browsing interface. Technical Report A32, Helsinki University of Technology, Laboratory of Computer and Information Science, Espoo, Finland.

Kaski, S., Honkela, T., Lagus, K., and Kohonen, T. (1998a). WEBSOM—self-organizing maps of document collections. *Neurocomputing*, 21:101–117.

Kaski, S., Kangas, J., and Kohonen, T. (1998b). Bibliography of self-organizing map (SOM) papers: 1981–1997. *Neural Computing Surveys*, 1(3&4):1–176. Available in electronic form at http://www.icsi.berkeley.edu/∼jagota/NCS/: Vol 1, pp. 102–350.

Kohonen, T. (1982). Self-organizing formation of topologically correct feature maps. *Biological Cybernetics*, 43(1):59–69.

Kohonen, T. (1995). *Self-Organizing Maps*, volume 30 of *Springer Series in Information Sciences*. Springer, Berlin, second, extended edition, 1997 edition.

Kohonen, T., Kaski, S., Lagus, K., Salojärvi, J., Paatero, V., and Saarela, A. (2000). Organization of a massive document collection. *IEEE Transactions on Neural Networks, Special Issue on Neural Networks for Data Mining and Knowledge Discovery*, 11(3):574–585.

Koskenniemi, K. (1983). *Two-Level Morphology: A General Computational Model for Word-Form Recognition and Production*. PhD thesis, University of Helsinki, Department of General Linguistics.

Lagus, K., Honkela, T., Kaski, S., and Kohonen, T. (1999). WEBSOM for textual data mining. *Artificial Intelligence Review*, 13(5/6):345–364.

Salton, G. and Buckley, C. (1987). Term weighting approaches in automatic text retrieval. Technical Report 87-881, Cornell University, Department of Computer Science, Ithaca, NY.

Salton, G. and McGill, M. J. (1983). *Introduction to modern information retrieval*. McGraw-Hill, New York.

Salton, G., Wong, A., and Yang, C. S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620.

Voorhees, E. M. and Harman, D. K. (2000). Appendix: Evaluation techniques and measures. In *Proceedings of The Eighth Text REtrieval Conference (TREC 8)*. NIST. http://trec.nist.gov/pubs/trec8/t8_proceedings.html.