DOCUMENT INDEXING USING INDEPENDENT TOPIC EXTRACTION

Yu-Hwan Kim and *Byoung-Tak Zhang*

School of Computer Science and Engineering
Seoul National University
Seoul 151-742, Korea
{yhkim,btzhang}@bi.snu.ac.kr

ABSTRACT

Text retrieval involves finding relevant information from a collection of documents given the user's information need. Traditional information retrieval systems represent a document as a vector of words, where each component can be a simple word count or follows a more sophisticated weighting scheme. The composed matrix is usually sparse and some words are highly correlated with each other. Another representation scheme is focused on the underlying topics which is usually obtained by a variety of dimension reduction techniques, where each document can be represented as a vector of topic intensity. In this paper, We proposed a novel indexing technique based on independent component analysis. From the experiments performed on AP news articles, the performance improvements is significant when the topic of query is closely related to the topics which is extracted from the ICA.

1. INTRODUCTION

Information retrieval involves retrieving relevant documents from a collection of free-form document given user's information request. In order to deal with free-form documents, we should represent documents in a standardized form. The most simple method is representing a document as a vector of word count. More sophisticated weighting scheme which are commonly used in the information retrieval community is based on *tf* · *idf*, where *tf* denotes word count that appears in the document and idf denotes inverse document frequency where document frequency is the number of documents which contain the word. It has the effect of setting large value for parsimonious words and relatively small value for commonly used words. In other words, idf has the effect of focusing on unique features. As a document is represented as a vector, a collection of documents can be represented as a matrix. Here, the number of rows corresponds to the size of vocabulary and the number of columns to the size of documents. Clearly, this term-document matrix is highly sparse, only a few having non-zero values. In addition, some words have the same sense (synonyms) or some words may have multiple senses in different contexts (polysems). Analysing such ambiguities is yet another issue in the information retrieval community.

Another method is topic-based document representation. Each document is assumed to be composed of a few topics and each topic has its own vocabulary. The final document is assumed to be generated by a linear mixture of words chosen from topics. In this scheme, we expect that the exact sense of an ambiguous word (synonyms and polysems) can be captured by analyzing topics which generated the word.

Though most documents seem to have such an implicit structure, mining underlying structures from free-form documents is not easy. With the advancement of several feature extraction and dimension reduction techniques, there have been proposed a variety of topic extraction algorithms [1, 2, 3]. Once underlying topics are found, these can be used for indexing, summarizing or visualizing a document [1]. For example, extracted topics can be used for making browsing interfaces for exploring information collections, which can be viewed as a method to allow humans to obtain a sense of the type of information stored in a corpus in order to perform their own categorization on the corpus.

In this paper, we try to find a novel topic-based indexing scheme by extracting independent features on the corpus using independent component analysis (ICA). ICA is an emerging technique for extracting statistically independent signals from mixed signals. It becomes a standard technology in blind source separation and deconvolution. In speech separation, independent components are equivalent to source signals [4]. It becomes more vague when it comes to image. Surprisingly, Bell and Sejnowski suggest that independent components of natural scenes are edge filters [5]. Likewise, it can be reasonablly inferred that the independent component of a document may be topics dealt in the document. Viewed from this point of view, a document can be considered as a linear mixture of several topics. However, the importance of topics should be measured appropriately, to be applied in a practical information retrieval problem. In other words, an important topic, e.g. topic from first principal component, should gain more weights because it is more useful in discriminating a set of documents from another. On the other hand, sparse topics should be given lower weights by the same reason. Conceptually both of them are the same as the idf measure acting as global weights.

In order to evaluate the performance of the independent component indexing scheme, we compared our methods against principal component analysis which is closely related to the latent semantic indexing technique [2] for the large scale corpus. Tested on a TREC-7 AP datasets, this approach showed significant improvements over principal component analysis when the topic of query is closely related to the extracted topics from independent component analysis.

The paper is organized as follows. In Section 2, we briefly discuss related work on topic extraction methods used in information retrieval community. Section 3 describes the ICA-based topic extractor. In Section 4, we report experimental results on TREC-7 and TREC-8 filtering datasets. Finally, Section 5 concludes this study.

2. TOPIC-BASED INDEXING TECHNIQUES

2.1. Latent Semantic Indexing

Latent semantic indexing (LSI) is a standard technique for reducing high dimensional term-based representation into small dimentional spaces where each dimension or factor is uncorrelated each other. LSI is based on singular value decomposition (SVD) which is defined as

$$A = U \Sigma V^T,$$

where Σ is a diagonal matrix composed of nonzero eigenvalues of AA^T or A^TA, and U and V are the orthonormal eigenvectors associated with the r nonzero eigenvalues of AA^T and A^TA, respectively. In some sense, the SVD can be viewed as a technique for deriving a set of uncorrelated indexing variables or factors, whereby each term and document is represented by a vector in $k \leq r$-space using columns of the U and V. It is worth noting that a document is represented by a vector in topic space inversely scaled by the locality of topics in LSI. This has the effect of emphasizing topics occurring in most of the documents than topics occurring in only a few documents. LSI have strong connection with principal component analysis (PCA). PCA is a standard technique for reducing dimensions by projecting from high-dimensional space into low-dimensional uncorrelated regions. PCA can be implemented based on SVD of mean-removed signals. Therefore, if we ignore scaling in LSI, documents in the latent semantic and principal compoenent space becomes almost the same.

2.2. Principal Component Analysis

Principal component analysis, or PCA [6] is widely used in signal and image processing. PCA can be defined using a recursive formulation.

$$\mathbf{w}_1 = \arg\max_{||w||=1} E(\mathbf{w}^T\mathbf{x})^2 \tag{1}$$

Thus the first principal component (PC) is the projection onto the direction in which the variance of the projection is maximized. Given the $(k-1)$th PC, the kth PC is determined as follows:

$$\mathbf{w}_k = \arg\max_{||w||=1} E[\mathbf{w}^T(\mathbf{x} - \sum_{i=1}^{k-1} \mathbf{w}_i \mathbf{w}_i^T \mathbf{x})]^2. \tag{2}$$

PCA can be easily implemented by a Hebbian-type neural networks [6]:

$$y_j = \sum_{i=1}^{m} w_{ji}(n) x_i(n) \tag{3}$$

$$\Delta w_{ji}(n) = \eta \left[y_j(n) x_i(n) - y_j(n) \sum_{k=1}^{j} w_{ki}(n) y_k(n) \right] \tag{4}$$

where m is the dimension of vector $\mathbf{x}$, n is the number of iterations. n is first set to 1. Then we iterate the above formula with increasing n while a stopping condition is met.

As is well known, principal component analysis can extract salient features from images. Likewise, we can expect that it may extract document topics applied to text document.

Table 1 visualizes the first five axes which have maximal variances. Here, we used AP news articles for the year 1988 which consist of 79919 documents. After removing stop-list (which occurs too frequently in the document like a, the, etc.) and applying Porter's stemming algorithm, we extracted about 20000 words for the indexing terms according to the increasing document frequency. Then we continued to iterating until 64 principal components are sufficiently stabilized.

From this result, one can easily find that the first and most important topic in 1988 was related to a presidential election. Note that the words with negative weights and words with positive weights are each highly correlated. On the other hand, the connection between the words with positive weights and negative weights looks more vague. For example, the words with positive weights are about presidential election while the words with negative weights are about economics in the second topic. Likewise, in the fifth topics, the words with positive weights deal with issues involving soviet union while the words with negative weights are about conflicts around the world. Therefore, we can conclude that each topic is decomposed into two subtopics, where all the subtopics are closely related with each other in a broad category.

Table 1. The first five principal components extracted using the PCA algorithm from the AP news articles. Odd columns show a few important words which are chosen according to the increasing weights. Even columns show corresponding weights.

1		2		3		4		5	
bush	-0.2975	cent	-0.0519	dukaki	-0.2750	cent	-0.3323	million	-0.3409
presid	-0.1520	polic	-0.0333	bush	-0.1966	million	-0.2800	compani	-0.2261
democrat	-0.1317	stock	-0.0329	jackson	-0.1517	market	-0.2298	court	-0.1332
campaign	-0.1201	market	-0.0239	democrat	-0.1151	price	-0.2168	feder	-0.1274
republican	-0.1077	compani	-0.0218	campaign	-0.1004	trade	-0.2023	billion	-0.1046
...									
price	0.0531	democrat	0.2242	report	0.1484	fire	0.0537	higher	0.0778
month	0.0645	state	0.2454	state	0.1795	peopl	0.0674	futur	0.1593
report	0.0663	bush	0.2830	offici	0.2056	jackson	0.0758	lower	0.1664
rate	0.0914	dukaki	0.3895	soviet	0.2831	kill	0.0873	soviet	0.1687
percent	0.8363	percent	0.4740	govern	0.2852	polic	0.1437	cent	0.4944

3. INDEPENDENT COMPONENT INDEXING

3.1. Independent Component Analysis

Independent component analysis is an emerging technique for extracting statistically indepedent components from data [7, 8]. As is well known, statistical independence can be defined as follows:

$$p(\mathbf{s}) = \prod_{i=1}^{M} p_i(s_i) \tag{5}$$

Assume that the observed vector $\mathbf{x}$ is generated by the linear mixing of the independent source signal $\mathbf{s}$:

$$\mathbf{x} = \mathbf{A}\mathbf{s}, \tag{6}$$

where only $\mathbf{x}$ is observable while other variables such as A and s should be estimated. Instead of the mixing matrix $\mathbf{A}$, ICA estimates its inverse $\mathbf{W}$, so that the outputs of the network $\mathbf{s}$ is statistically independent

$$\mathbf{s} = \mathbf{W}\mathbf{x}. \tag{7}$$

Sejnowski et al. [4] established ICA algorithms on the basis of entropy maximization to estimate super-gaussian sources.

$$\Delta\mathbf{W} \propto \left[\mathbf{W}^T\right]^{-1} - 2\tanh(\mathbf{W}\mathbf{u})\mathbf{u}^T, \tag{8}$$

where the $\tanh$ function is used for estimating super-gaussian independent compoenents. Other function should be used for sub-Gaussian sources.

The convergence of the gradient method can be greatly improved by the introduction of natural gradient algorithm proposed by Amari [9]. In the case of ICA, applying natural gradient has the effect of multiplying $\mathbf{W^T W}$ at the end of the Equation 8.

$$\Delta\mathbf{W} \propto (\mathbf{I} - 2\tanh(\mathbf{u})\mathbf{u}^T)\mathbf{W} \tag{9}$$

More recently, Lee et al. [10] proposed an extended ICA algorithm for both super-gaussian and sub-gaussian source.

$$\Delta\mathbf{W} \propto \left[\mathbf{I} - \mathbf{K}\tanh(\mathbf{u})\mathbf{u}^T - \mathbf{u}\mathbf{u}^T\right]\mathbf{W}, \tag{10}$$

where $\mathbf{K}$ is a diagonal matrix and $k_i = 1$ if the estimated source u_i is super-Gaussian and $k_i = -1$ if sub-Gaussian.

Independently, Hyvarinen [11] proposed a fixed-point algorithm for separating mixed signals of both super-gaussian and sub-gaussian sources. Though this algorithm needs preliminary sphering, it showed fast convergence. The relationships among various models such as factor analysis, PCA, and ICA become clear by the work of Attias [12].

One of the drawbacks of ICA is that we cannot determine the variances of the independent components because any scalar multiplier in one of the sources s_i can be mitigated by dividing the corresponding column $\mathbf{a}_i$ by the same scalar [13]. Though this ambiguity is insignificant in most applications, we can no longer set global weights naturally as in the LSI. As is well known in the information retrieval community, when we determine weights on the basis of features, we multiply the local weights and global weights each representing the number of features occurring a specific document and the spreadness of features across all the documents.

Most ICA algorithms assume equal numbers of sources and sensors. When there are more sensors than sources we should reduce dimensions using PCA. Accordingly, we first reduce dimensions into 64 from about 20000. Then, we performed extended ICA algorithms producing 64 independent components. In Table 2, we illustrate several independent

Table 2. The first five independent components extracted using extended ICA algorithm. In each sub-table, the odd first columns show a few important words which are chosen according to the increasing weights. The even columns show corresponding weights.

1		2		3		4		5	
state	-0.0892	compani	-0.0662	dukaki	-0.3438	million	-0.3972	million	-0.4333
dukaki	-0.0812	million	-0.0503	jackson	-0.2626	cent	-0.3594	company	-0.2819
campaign	-0.0694	cent	-0.0472	bush	-0.1396	company	-0.2650	share	-0.1406
bush	-0.0513	trade	-0.0436	cent	-0.1214	market	-0.1690	offer	-0.1021
presid	-0.0351	market	-0.0412	campaign	-0.0881	bank	-0.1672	feder	-0.0880
...									
precinct	0.0520	vote	0.1076	percent	0.1486	citi	0.0843	fire	0.1032
rate	0.0536	bush	0.1649	fire	0.1761	percent	0.0882	higher	0.1224
voter	0.0603	state	0.2393	unit	0.1796	kill	0.0949	futur	0.1611
month	0.0696	dukaki	0.3020	soviet	0.2457	polic	0.1110	lower	0.1658
percent	0.9387	percent	0.7990	state	0.3551	jackson	0.1166	cent	0.4968

components extracted from AP news articles. The weights shown in the figure is normalized to one so as to make the direct comparison with PCA possible.

The results look similar to that of PCA. However, most weights are concentrated on a small set of words compared with PCA. It means that these words play a crucial role in the topic-based representation. For example, documents dealing with topics discovered by ICA might have higher norms. It has the effect of spreading the transformed vector in a broad area allowing discrimination of documents easier. On the other hand, a document dealing with a topic undiscovered by the ICA might be transformed into a representation with a relatively small norm. It makes classifying documents difficult since we should discriminate documents in a rather small area. We will discuss this topic again when we deal with information retrieval problems in the next section.

3.2. Independent Component Indexing

From Tables 1 and 2, ICA is assumed to be able to extract more salient features than PCA. However, it is still incomplete to be used in information retrieval problems. The success of *tf* · *idf* and latent semantic indexing is partially based on the intelligent use of term or topic locality. Likewise, adding global weights like lantent semantic indexing will expect more performance improvement. Following the scheme of latent semantic indexing, global weights are measured by the root of the variances of an independent components. The final weight of the jth feature in the document d_i is as follows:

$$w_{ij} = \text{local weight}_{ij} * \text{global weight}_j \qquad (11)$$

$$= d_{ij} * \sqrt{\sum_i (d_{ij} - \bar{d}_{ij})^2}, \qquad (12)$$

Table 3. General domains of query statements

Topic No.	Domain
1–5, 8	International Economics
6	International Finance
7	U. S. Economics
9	U.S. Politics
10–11	Science & Technology
12	Environment

where d_{ij} is the weight of the jth feature in the document d_i. Since estimating the variance of an independent component is impossible, we restrict our ICA algorithm so that the length of each row is unity. Note that this is a heuristic scheme and another weighting scheme can be used instead.

Now, the main task is retrieving documents which are highly ranked accroding to a similarity measure given a user's query statement. The most commonly used similarity measure is based on cosine similarity ($s_{d_i q}$) which is defined as follows:

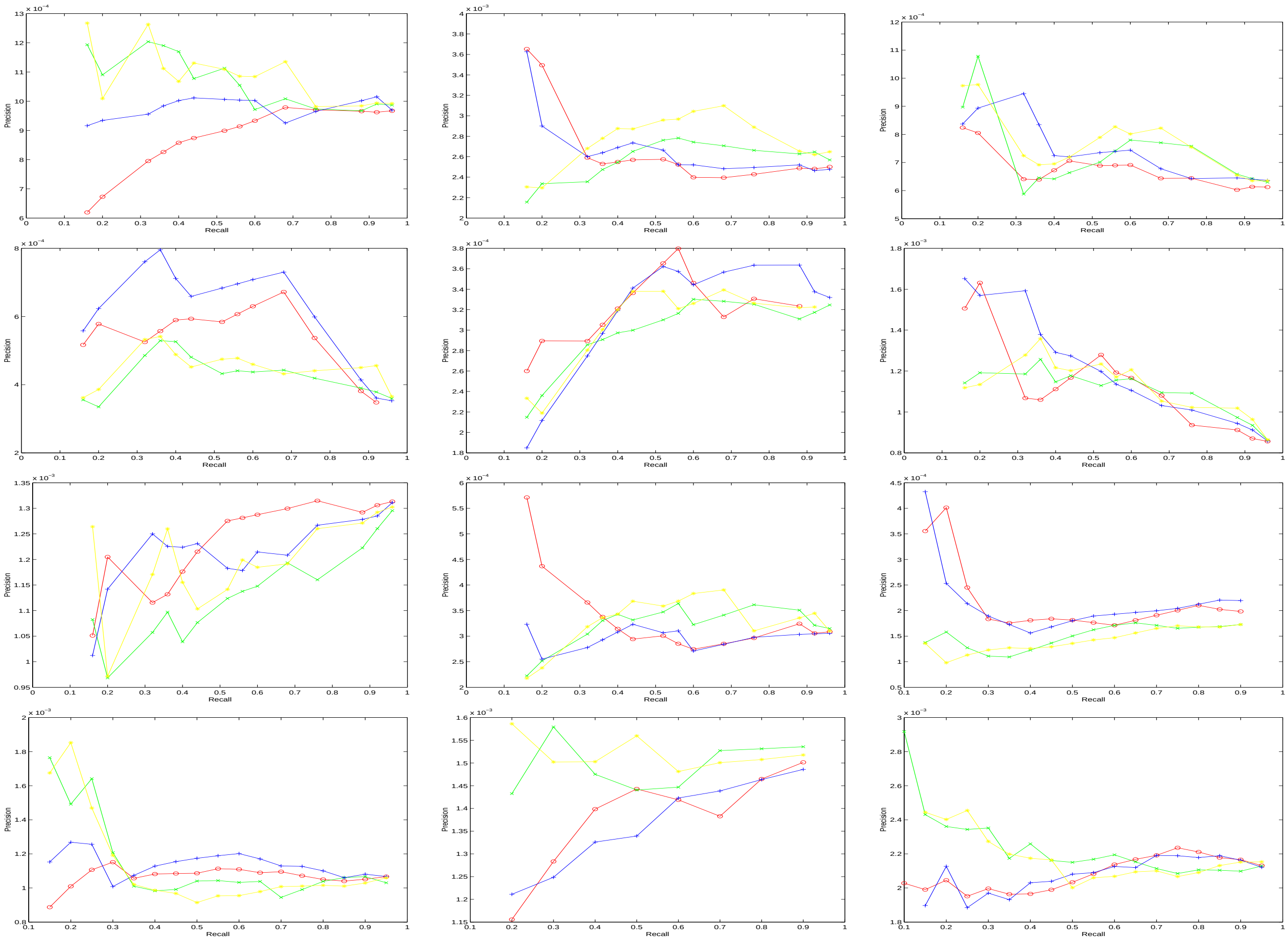
$$s_{d_i q} = \frac{\sum_j d_{ij} q_j}{\sqrt{\sum_j d_{ij}^2}\sqrt{q_j^2}}, \qquad (13)$$

where q_j is the weight of jth feature of query q. In the ordinary information retrieval context, d_{ij} and q_j can be $tf \cdot idf$ while it is 11 in the independent component based representation

4. EXPERIMENTS

We used articles from AP News for the years 1988, and 12 query statements which are obtainable from TREC dataset

Fig. 1. The performance of various topic based indexing techniques. AP news articles and TREC topics 1–8 was used for evaluation. Dark line with circle, dark line plus, light line with x-mark, light line with star shows the performance of ICA, ICA with global weights, PCA, and PCA with global weights respectively. Linear interpolated precision-recall graphs are shown for 12 query statements. The results are shown in the first row for the query 1–3, second row for the query 4–6, third row for the query 7–9, and fourth row for the query 10–12.



[14]. Table 3 summarizes general domains of query statements.

We stemmed the documents using the Porter's algorithm, and removed words from the stop-list and common short words. Then, we removed documents which had fewer than 20 terms, and extracted about 20000 terms according to the increasing document frequency. We used precision and recall for evaluating the performance of the system which is a de facto standard in the information retrieval community:

$$precision = \frac{\text{\# of retrieved relevant documents}}{\text{\# of retrieved documents}} \tag{14}$$

$$recall = \frac{\text{\# of retrieved relevant documents}}{\text{\# of relevant documents}} \tag{15}$$

We evaluated the performance of various topic-based indexing schemes, i.e., PCA, PCA with global weights, ICA, and ICA with global weights. PCA with global weights are similar to latent semantic indexing as shown in Section 2.1. Figure 1 shows the performance of each scheme. It is clear that the performance of ICA is promising showing good performance except one case. Additionally, ICA with global weights shows similar or better performance than ICA based indexing scheme. It is because the cosine measure used is based on a vector space model (VSM) where features are asssumed to be mutually independent [15]. However, ICA with global weights showed only moderate performance when tested with the 1st (upper-left pane), 10–12th (panes shown in the 4th row) query statements. Among those, 10-12th query statements are about science & technology. As is shown above, the independent components extracted are mostly about ecnomics, politics, and a variety of conflictions around the world. Accordingly, the performance deterioration comes from the lack of explicit independent component representing science & technology. The first query statement which deals with *Antitrust Cases Pending* which does not seem to be a major topic. Actually, the norm of the word ‘antitrust’ is only 0.0103 while the word ‘acquisition’ which can be found in the second query statement marks 0.0406. From these observations, we can infer that independent component indexing improves retrieval performance only when querying about the topics which are extracted from ICA.

5. CONCLUSIONS

We proposed a novel ICA-based indexing scheme and evaluated its performance for a large real-world dataset composed of about 80000 documents. The performance can be improved further by adopting a global weight factor similar to idf factor used frequently in the information retrieval community. It is interesting to note that the performance improvements are observed only when the topic of a query is closely related with the topics extracted from the ICA algorithm. However, our ICA-based indexing schmeme shows only marginal performance when the topic of a query is different from the extracted topics. Based on this result, our future work will be proposing a more general purpose topic-based indexing algorithms.

Acknowledgements

This research was supported by Korea Ministry of Information and Telecommunication under Grant 98-199 through IITA, and by the BK21-IT Program.

6. REFERENCES

[1] T. Hofmann, “Probabilistic latent semantic indexing,” in *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1999, pp. 50–57.

[2] S. T. Dumais, “Latent semantic indexing (LSI): Trec-3 report,” in *Proceedings of the Text REtrieval Conference (TREC-3)*, 1995, pp. 219–230.

[3] R. Swan and J. Allan, “On-line new event detection and tracking,” in *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2000, pp. 76–84.

[4] A. J. Bell and T. J. Sejnowski, “An information maximization approach to blind separation and blind deconvolution,” *Neural Computation*, vol. 7, no. 6, pp. 1129–1159, 1995.

[5] A. J. Bell and T. J. Sejnowski, “The ’independent components’ of natural scenes are edge filters,” *Vision Research*, vol. 37, no. 23, pp. 3327–3338, 1997.

[6] E. Oja, “A simplified neuron model as a principal component analyzer,” *Journal of Mathematical Biology*, vol. 15, pp. 267–273, 1982.

[7] P. Comon, “Independent component analysis - a new concept?,” *Signal Processing*, vol. 36, pp. 287–314, 1994.

[8] C. Jutten and J. Herault, “Blind separation of sources, part i: An adaptive algorithm based on neuromimetic architecture,” *Signal Processing*, vol. 24, pp. 1–10, 1991.

[9] S.-I. Amari, A. Cichocki, and H. H. Yang, “A new learning algorithm for blind source separation,” in *Advances in Neural Information Processing Systems*, 1996, pp. 757–763.

[10] T.-W. Lee, M. Girolami, and T. J. Sejnowski, “Independent component analysis using an extended infomax algorithm for mixed sub-gaussian and super-gaussian sources,” *Neural Computation*, vol. 11, no. 2, pp. 417–441, 1999.

[11] A. Hyvärinen, “Fast and robust fixed-point algorithm for independent component analysis,” *IEEE Trans. on Neural Networks*, vol. 10, no. 3, pp. 626–634, 1999.

[12] H. Attias, “Independent factor analysis,” *Neural Computation*, vol. 11, no. 4, pp. 803–851, 1999.

[13] A. Hyvärinen, “Survey on independent component analysis,” *Neural Computing Surveys*, vol. 2, pp. 94–128, 1999.

[14] E. Voorhees and D. Harman, “Overview of the seventh text retrieval conference (TREC-7),” in *The Seventh Text REtrieval Conference (TREC-7)*, 1998, pp. 1–24.

[15] G. Salton and M. J. McGill, *Introduction to modern information retrieval*, McGraw-Hill, 1983.