Information Retrieval Capable of Visualization and High Precision

Qing Ma[1,2] and **Kousuke Enomoto**[1]
[1]Ryukoku University / [2]NICT, Japan
qma@math.ryukoku.ac.jp

Masaki Murata and **Hitoshi Isahara**
NICT, Japan
{murata,isahara}@nict.go.jp

Abstract

We present a neural-network based self-organizing approach that enables visualization of the information retrieval while at the same time improving its precision. In computer experiments, two-dimensional documentary maps in which queries and documents were mapped in topological order according to their similarities were created. The ranking of the results retrieved using the maps was better than that of the results obtained using a conventional TFIDF method. Furthermore, the precision of the proposed method was much higher than that of the conventional TFIDF method when the process was focused on retrieving highly relevant documents, suggesting that the proposed method might be especially suited to information retrieval tasks in which precision is more critical than recall.

1 Introduction

Information retrieval (IR) has been studied since an earlier stage [e.g., (Menzel, 1966)] and several kinds of basic retrieval models have been proposed (Salton and Buckley, 1988) and a number of improved IR systems based on these models have been developed by adopting various NLP techniques [e.g., (Evans and Zhai, 1996; Mitra et al., 1997; Mandara, et al., 1998; Murata, et al., 2000)]. However, an epoch-making technique that surpasses the TFIDF weighted vector space model, the main approach to IR at present, has not yet been invented and IR is still relatively imprecise. There are also challenges presenting a large number of retrieval results to users in a visual and intelligible form.

Our aim is to develop a high-precision, visual IR system that consists of two phases. The first phase is carried out using conventional IR techniques in which a large number of related documents are gathered from newspapers or websites in response to a query. In the second phase the visualization of the retrieval results and picking are performed. The visualization process classifies the query and retrieval results and places them on a two-dimensional map in topological order according to the similarity between them. To improve the precision of the retrieval process, the picking process involves further selection of a small number of highly relevant documents based on the classification results produced by the visualization process.

This paper presents a new approach by using the self-organizing map (SOM) proposed by Kohonen (Kohonen, 1997) for this second IR phase[1]. To enable the second phase to be slotted into a practical IR system as described above, visual-

[1]There have been a number of studies of SOM on data mining and visualization [e.g., (Kohonen, et al., 2000)] since the WEBSOM was developed in 1996. To our knowledge, however, these works mainly focused on confirming the capabilities of SOM in the self-organization and/or in the visualization. In this study, we slot the SOM-based processing into a practical IR system that enables visualization of the IR while at the same time improving its precision. The another feature of our study differing from others is that we performed comparative studies with TFIDF-based IR methods, the major approach to IR in NLP field.

ization and picking should be carried out for a single query and set of related documents. In this paper, however, for the purpose of evaluating the proposed system, correct answer data, consisting of multiple queries and related documents as used in the 1999 IR contest, IREX (Murata, et al., 2000), was used. The procedure of the second IR-phase in this paper is therefore as follows. Given a set of queries and related documents, a documentary map is first automatically created through self-organization. This map provides visible and continuous retrieval results in which all queries and documents are placed in topological order according to their similarity[2]. The documentary map provides users with an easy method of finding documents related to their queries and also enables them to see the relationships between documents with regard to the same query, or even the relationships between documents across different queries. In addition, the documents related to a query can be ranked by simply calculating the Euclidean distances between the points of the queries and the points of the documents in the map and then choosing the N closest documents in ranked order as the retrieval results for each query. If a small N is set, then the retrieval results are limited to the most highly relevant documents, thus improving the retrieval precision.

Computer experiments showed that meaningful two-dimensional documentary maps could be created; The ranking of the results retrieved using the map was better than that of the results obtained using a conventional TFIDF method. Furthermore, the precision of the proposed method was much higher than that of the conventional TFIDF method when the retrieval process focused on retrieving the most highly relevant documents, which indicates that the proposed method might be particularly useful for picking the best documents, thus greatly improving the IR precision.

2 Self-organizing documentary maps and ranking related documents

A SOM can be visualized as a two-dimensional array of nodes on which a high-dimensional input vector can be mapped in an orderly manner through a learning process. After the learning, a meaningful nonlinear coordinate system for different input features is created over the network. This learning process is competitive and unsupervised and is called a self-organizing process.

Self-organizing documentary maps are ones in which given queries and all related documents in the collection are mapped in order of similarity, i.e., queries and documents with similar content are mapped to (or best-matched by) nodes that are topographically close to one another, and those with dissimilar content are mapped to nodes that are topographically far apart. Ranking is the procedure of ranking documents related to each query from the map by calculating the Euclidean distances between the points of the queries and the points of the documents in the map and choosing the N closest documents as the retrieval result.

2.1 Data

The queries are those used in a dry run of the 1999 IREX contest and the documents relating to the queries are original Japanese newspaper articles used in the contest as the correct answers. In this study, only nouns (including Japanese verbal nouns) were selected for use.

2.2 Data coding

Suppose we have a set of queries:

$$Q = \{Q\_i \quad (i = 1, \cdots, q)\}, \tag{1}$$

where q is the total number of queries, and a set of documents:

$$A = \{Ai\_j \quad (i = 1, \cdots, q, j = 1, \cdots, a_i)\}, \tag{2}$$

where a_i is the total number of documents related to Q_i. For simplicity, where there is no need to distinguish between queries and documents, we use the same term "documents" and the same notation D_i to represent either a query Q_i or a document Ai_j. That is, we define a new set

$$D = \{D_i \quad (i = 1, \cdots, d)\} = Q \bigcup A \tag{3}$$

which includes all queries and documents. Here, d is the total number of queries and documents, i.e.,

$$d = q + \sum_{i=1}^{q} a_i. \tag{4}$$

[2]For a specific query, other queries and documents in the map are considered to be irrelevant (i.e., documents unrelated to the query). This map is therefore equivalent to a map consisting of one query and related and unrelated documents, which will be adopted in the practical IR system that we aim to develop.

Each document, D_i, can then be defined by the set of nouns it contains as

$$D_i = \{noun_1^{(i)}, w_1^{(i)}, \cdots, noun_{n_i}^{(i)}, w_{n_i}^{(i)}\}, \quad (5)$$

where $noun_k^{(i)}$ ($k = 1, \cdots, n_i$) are all different nouns in the document D_i and $w_k^{(i)}$ is a weight representing the importance of $noun_k^{(i)}$ ($k = 1, \cdots, n_i$) in document D_i. The weights are computed by their **tf** or **tfidf** values. That is,

$$w_j^{(i)} = \mathrm{tf}_j^{(i)} \text{ or } \mathrm{tf}_j^{(i)}\mathrm{idf}_j. \quad (6)$$

In the case of using tf, the weights are normalized such that

$$w_1^{(i)} + \cdots + w_{n_i}^{(i)} = 1. \quad (7)$$

Also, when using the Japanese thesaurus, Bunrui Goi Hyou (The National Institute for Japanese Language, 1964) (BGH for short), synonymous nouns in the queries are added to the sets of nouns from the queries shown in Eq. (5) and their weights are set to be the same as those of the original nouns.

Suppose we have a correlative matrix whose element d_{ij} is some metric of correlation, or a similarity distance, between the documents D_i and D_j; i.e., the smaller the d_{ij}, the more similar the two documents. We can then code document D_i with the elements in the i-th row of the correlative matrix as

$$V(D_i) = [d_{i1}, d_{i2}, \cdots, d_{id}]^T. \quad (8)$$

The $V(D_i) \in \Re^d$ is the input to the SOM. Therefore, the method to compute the similarity distance d_{ij} is the key to creating the maps. Note that the individual d_{ij} of vector $V(D_i)$ only reflects the relationships between a pair of documents when they are considered independently. To establish the relationships between the document D_i and all other documents, representations such as vector $V(D_i)$ are required. Even if we have these high-dimensional vectors for all the documents, it is still difficult to establish their global relationships. We therefore need to use an SOM to reveal the relationships between these high-dimensional vectors and represent them two-dimensionally. In other words, the role of the SOM is merely to self-organize vectors; the quality of the maps created depends on the vectors provided.

In computing the similarity distance d_{ij} between documents, we take two factors into account: (1) the larger the number of common nouns in two documents, the more similar the two documents should be (i.e., the shorter the similarity distance); (2) the distance between any two queries should be based on their application to the IR processing; i.e., by considering the procedure used to rank the documents relating to each query from the map. For this reason, the document-similarity distance between queries should be set to the largest value. To satisfy these two factors, d_{ij} is calculated as follows:

$$d_{ij} = \begin{cases} 1 & \text{if both } D_i \text{ and } D_j \text{ are queries} \\ 1 - \frac{|C_{ij}|}{|D_i|+|D_j|-|C_{ij}|} & \text{not the case mentioned above and } i \neq j \\ 0, & \text{if i=j} \end{cases} \quad (9)$$

where $|D_i|$ and $|D_j|$ are values (the numbers of elements) of sets of documents D_i and D_j defined by Eq. (5) and $|C_{ij}|$ is the value of the intersection C_{ij} of the two sets D_i and D_j. $|C_{ij}|$ is therefore some metric of document similarity (the inverse of the similarity distance d_{ij}) between documents D_i and D_j which is normalized by $|D_i|+|D_j|-|C_{ij}|$. Before describing the methods for computing them, we first rewrite the definition of documents given by Eq. (5) for D_i and D_j as follows.

$$D_i = \{(c_1, w_{c1}^{(i)}, \cdots, c_l, w_{cl}^{(i)}), (n_1^{(i)}, w_1^{(i)}, \cdots, n_{m_i}^{(i)}, w_{m_i}^{(i)})\}, \quad (10)$$

and

$$D_j = \{(c_1, w_{c1}^{(j)}, \cdots, c_l, w_{cl}^{(j)}), (n_1^{(j)}, w_1^{(j)}, \cdots, n_{m_j}^{(j)}, w_{m_j}^{(j)})\}, \quad (11)$$

where c_k ($k = 1, \cdots, l$) are the common nouns of documents D_i and D_j and $n_k^{(i)}$ ($k = 1, \cdots, m_i$) and $n_k^{(j)}$ ($k = 1, \cdots, m_j$) are nouns of documents D_i and D_j which differ from each other. By comparing Eq. (5) and Eqs. (10) and (11), we know

that $l + m_i + m_j = n_i + n_j$. Thus, $|D_i|$ (or $|D_j|$) of Eq. (9) can be calculated as follows.

$$|D_i| = \sum_{k=1}^{l} w_{ck}^{(i)} + \sum_{k=1}^{m_i} w_k^{(i)}. \quad (12)$$

For calculating $|C_{ij}|$, on the other hand, since the weights (of either common or different nouns) generally differ between two documents, we devised four methods which are expressed as follows.

Method A:

$$|C_{ij}| = \sum_{k=1}^{l} \max(w_{ck}^{(i)}, w_{ck}^{(j)}). \quad (13)$$

Method B:

$$|C_{ij}| = \sum_{k=1}^{l} \frac{w_{ck}^{(i)} + w_{ck}^{(j)}}{2}. \quad (14)$$

Method C:

$$|C_{ij}| = \begin{cases} \sum_{k=1}^{l} \max(w_{ck}^{(i)}, w_{ck}^{(j)}) & \text{if one is a query and the other is a document} \\ \sum_{k=1}^{l} \frac{w_{ck}^{(i)} + w_{ck}^{(j)}}{2}. & \text{if both are documents} \end{cases} \quad (15)$$

Method D:

$$|C_{ij}| = \begin{cases} \sum_{k=1}^{l} \max(w_{ck}^{(i)}, w_{ck}^{(j)}) & \text{if one is a query and the other is a document} \\ \sum_{k=1}^{l} \min(w_{ck}^{(i)}, w_{ck}^{(j)}). & \text{if both are documents} \end{cases} \quad (16)$$

Note that we need not consider the case where both are queries for calculating $|C_{ij}|$ because this has been considered independently as shown by Eq. (9).

3 Experimental Results

3.1 Data

Six queries Q_i ($i = 1, \cdots, q$, $q = 6$) and 433 documents Ai_j ($i = 1, \cdots, q$, $q = 6$, $j = 1, \cdots, a_i$ and $\sum_{i=1}^{q} a_i = 433$) used in the dry run of the 1999 IREX contest were used for our experiments. The distribution of these documents, i.e., the number a_i ($i = 1, \cdots, q$, $q = 6$) of documents related to each query, is shown in Table 1.

Table 1: Distribution of documents used in the experiments

a_1	a_2	a_3	a_4	a_5	a_6	$\sum_{i=1}^{6} a_i$
80	89	42	108	49	65	433

It should be noted that since the proposed IR approach will be slotted into a practical IR system in the second phase in which a small number (say below 1,000, or even below 500) of the related documents should have been collected, this experimental scale is definitely a practical one.

3.2 SOM

We used a SOM of a 40×40 two-dimensional array. Since the total number d of queries and documents to be mapped was 439, i.e., $d = q + \sum_{i=1}^{6} a_i = 439$, the number of dimensions of input n was 439. In the ordering phase, the number of learning steps T was set at 10,000, the initial value of the learning rate $\alpha(0)$ at 0.1, and the initial radius of the neighborhood $\sigma(0)$ at 30. In the fine adjustment phase, T was set at 15,000, $\alpha(0)$ at 0.01, and $\sigma(0)$ at 5. The initial reference vectors $\mathbf{m_i}(\mathbf{0})$ consisted of random values between 0 and 1.0.

3.3 Results

We first performed a preliminary experiment and analysis to determine which of the four methods was the optimal one for calculating $|C_{ij}|$ shown in Eqs. (13)-(16). Table 2 shows the IR precision, i.e., the precision of the ranking results obtained from the self-organized documentary maps created using the four methods. The IR precision was calculated by follows.

$$P = \frac{1}{q} \sum_{i=1}^{q} \frac{\#\text{related to } Q\_i \text{ in the retrieved } a_i \text{ documents}}{a_i}, \quad (17)$$

where q is the total number of queries, $\#$ means number, and a_i is the total number of documents related to Q_i as shown in Table 1.

In the case of using tf values as weights of nouns, method B obviously did not work. Al-

Table 2: IR precision for the four methods for calculating $|C_{ij}|$

Weight	Method A	Method B	Method C	Method D
tf	0.33	0.20	0.41	**0.45**
tfidf	0.85	0.76	**0.91**	0.78

though the similarity between queries was mandatorily set to the largest value, all six queries were mapped in almost the same position, thus producing the poorest result. We consider the reason for this was as follows. In general, the number of words in a query is much smaller than the number of words in the documents, and the number of queries is much smaller than the number of documents collected. As described in section 2, each query was defined by a vector consisting of all similarities between the query and five other queries and all documents in the collection. We think that using the average weights of words appearing in the queries and documents to calculate the similarities between queries and documents, as in method B, tends to produce similar vectors for the queries. All of these query vectors are then mapped to almost the same position. With coding method A, because the larger of the two weights of a query and a document is used, the same problem could also arise in practice. There were no essential differences between coding methods C and D, which were almost equally precise. Neither of these methods have the shortcomings described above for methods A and B. However, when tfidf values were used as the weights of the nouns, even methods A and B worked quite well. Therefore, if we use tfidf values as the weights of the nouns, we may use either of the four methods. Based on this analysis and the preliminary experimental result that method C and D had highest precisions in the cases of using tf and tfidf values as weights of the nouns, respectively, we used methods C and D for calculating $|C_{ij}|$ in all the remaining experiments.

Table 3 shows the IR precision obtained using various methods. From this table we can see that the proposed method in the case of SOM (w=tfidf, C), i.e., using method C for calculating $|C_{ij}|$, using tfidf values as the weights of nouns, and not using the Japanese thesaurus (BGH), in the case of SOM (w=tfidf, D), i.e., using method D, using tfidf values, and not using the BGH, and in

Table 3: IR precision obtained using various methods

TFIDF	TFIDF (BGH)	SOM (w=tf, D)	**SOM (w= tfidf, C)**	SOM (w= tfidf, C, BGH)	SOM (w= tfidf, D)	SOM (w= tfidf, D, BGH)
0.67	0.75	0.45	**0.91**	0.77	0.78	0.73

Table 4: IR precision for top N related documents

N	TFIDF	TFIDF (BGH)	SOM (w=tf, D)	**SOM (w= tfidf, C)**	SOM (w= tfidf, C, BGH)	SOM (w= tfidf, D)	SOM (w= tfidf, D, BGH)
10	0.83	0.88	0.75	**1.0**	0.97	1.0	0.97
20	0.79	0.86	0.68	**0.99**	0.95	0.98	0.97
30	0.73	0.84	0.62	**0.99**	0.94	0.97	0.91
40	0.71	0.82	0.58	**0.98**	0.90	0.97	0.87

the case of SOM (w=tfidf, C, BGH), i.e., using method C, using tfidf values, and using the BGH produced the highest, second highest, and third highest precision, respectively, of all the methods including the conventional TFIDF method. When the BGH was used, however, the IR precision of the proposed method dropped inversely, whereas that of the conventional TFIDF improved. The lower precision of the proposed method when using BGH might be due to the calculation of the denominator of Eq. (9); this will be investigated in future study.

Table 4 shows the IR precision obtained using various methods when the retrieval process is focused on the top N related documents. From this table we can see that the IR precision of the proposed method, no matter whether the BGH was used or not, or whether method C or D was used for calculating $|C_{ij}|$, was much higher than that of the conventional TFIDF method when the process was focused on retrieving the most relevant documents. This result demonstrated that the proposed method might be especially useful for picking highly relevant documents, thus greatly improving the precision of IR.

Figure 1 shows the left-top area of a self-organized documentary map obtained using the proposed method in the case of SOM (w=tfidf, D)[3]. From this map, we can see that query Q_4

[3]Note that the map obtained using the proposed method in the case of SOM (w=tfidf, C), which had the highest IR precision, was better than this.

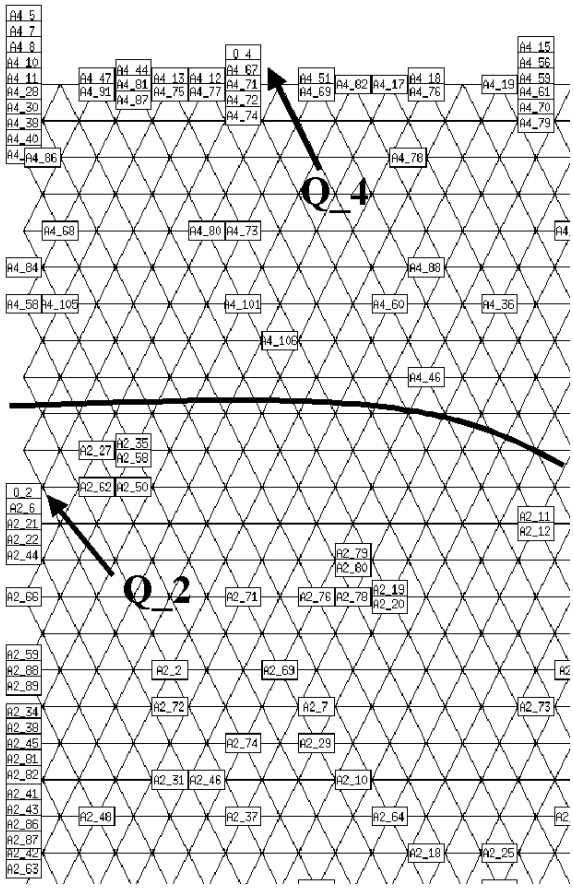

Figure 1: Left-top area of self-organized documentary map

and its related documents $A4\_*$ (where * denotes an Arabic numeral), Q_2 and its related documents $A2\_*$ were mapped in positions near each other. Similar results were obtained for the other queries which were not mapped in the area of the figure. This map provides visible and continuous retrieval results in which all queries and documents are placed in topological order according to their similarities. The map provides an easy way of finding documents related to queries and also shows the relationships between documents with regard to the same query and even the relationships between documents across different queries.

Finally, it should be noted that each map that consists of 400 to 500 documents was obtained in 10 minutes by using a personal computer with a 3GHZ CPU of Pentium 4.

4 Conclusion

This paper described a neural-network based self-organizing approach that enables information retrieval to be visualized while improving its precision. This approach has a practical use by slotting it into a practical IR system as the second-phase processor. Computer experiments of practical scale showed that two-dimensional documentary maps in which queries and documents are mapped in topological order according to their similarities can be created and that the ranking of the results retrieved using the created maps is better than that produced using a conventional TFIDF method. Furthermore, the precision of the proposed method was much higher than that of the conventional TFIDF method when the process was focused on retrieving the most relevant documents, suggesting that the proposed method might be especially suited to information retrieval tasks in which precision is more important than recall.

In future work, we first plan to re-confirm the effectiveness of using the BGH and to further improve the IR accuracy of the proposed method. We will then begin developing a practical IR system capable of visualization and high precision using a two-phase IR procedure. In the first phase, a large number of related documents are gathered from newspapers or websites in response to a query presented using conventional IR; the second phase involves visualization of the retrieval results and picking the most relevant results.

References

H. Menzel. 1966. Information needs and uses in science and technology. Annual Review of Information Science and Technology, 1, pp. 41-69.

G. Salton and C. Buckley. 1988. Term-weighting approaches in automatic text retrieval. Information Processing & Management, 24(5), pp. 513-523.

D. A. Evans and C. Zhai. 1996. Noun-phrase analysis in unrestricted text for information retrieval. ACL'96, pp. 17-24.

M. Mitra, C. Buckley, A. Singhal, and C. Cardie, C. 1997. An analysis of statistical and syntactic phrases. RIAO'97, pp. 200-214.

R. Mandara, T. Tokunana, and H. Tanaka 1998. The use of WordNet in information retrieval. COLING-ACL'98 Workshop: Usage of WordNet in Natural Language Processing Systems, pp. 31-37.

M. Murata, Q. Ma, K. Uchimoto, H. Ozaku, M. Uchiyama, and H. Hitoshi 2000. Japanese probabilistic information retrieval using location and category information. IRAL'2000.

T. Kohonen 1997. Self-organizing maps. Springer, 2nd Edition.

T. Kohonen, S. Kaski, K. Lagus, J. Salojarrvi, J. Honkela, V. Paatero, and A. Saarela. 2000. Self Organization of a Massive Document Collection. IEEE Trans. Neural Networks, 11, 3, pp. 574-585.

The National Institute for Japanese Language. 1964. Bunrui Goi Hyou (Japanese Thesaurus). Dainippon-tosho.