Single Language Information Retrieval at NTCIR-5

Qing Ma, Koichi Nakao, Kousuke Enomoto
Ryukoku University
Otsu 520-2194, Japan
qma@math.ryukoku.ac.jp

Masaki Murata
National Institute of Information and Communications Technology
Kyoto 619-0289, Japan
murata@nict.go.jp

Abstract

We participated in the Single Language (Japanese) Information Retrieval (IR), or SLIR, subtask of the Cross-Lingual Information Retrieval (CLIR) task at NTCIR-5 to verify the practical effectiveness of a two-phase IR system that we proposed for visualizing IR while at the same time improving its precision. Although the proposed system performed well in relatively small computer experiments, the results of the NTCIR-5 task were not as good as we expected, revealing the need for more work in adaptating our system for large-scale practical IR tasks. This paper describes the details of our systems submitted to the NTCIR-5 task and their experimental results, and proposes three solutions to the scaling problem that allow the proposed system to be effectively adapted for practical IR tasks. Very beginning additional experiments based on the solutions showed that our system has a comparable performance when using automatic method, and possibly a much higher precision when using a semi-automatic method, compared to the conventional TFIDF-based method.

1 Introduction

We previously proposed a two-phase information retrieval (IR) system aimed at developing a high-precision, visual IR system [1]. The first phase is carried out using conventional TFIDF-based techniques, in which a large number of relevant documents are gathered from newspapers or websites in response to a query. In the second phase, the visualization of the retrieval results and the 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. For this second IR phase, we proposed a new approach using the self-organizing map (SOM) proposed by Kohonen [2].

Relatively small computer experiments, in which the correct answers of the dry run of the 1999 IREX contest [3] were used, have 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-based method. Furthermore, the precision of the proposed method was much higher than that of the conventional TFIDF-based method when the retrieval process focused on retrieving the most highly relevant documents, which indicates the proposed method might be particularly useful for picking the most accurate documents, thus greatly improving IR precision.

To verify the practical effectiveness of our two-phase IR system, we participated in the Single Language (Japanese) IR (SLIR) subtask of the Cross-Lingual Information Retrieval (CLIR) task at NTCIR-5. Regrettably, the results of this task were not as good as we expected, and more work adapting our system for large-scale IR tasks is required. This paper describes the details of our systems submitted to the NTCIR-5 task and their experimental results, and proposes three solutions for scaling problem that allow the proposed system to be effectively adapted for practical IR tasks. Very beginning additional experiments based on the solutions shows that our system has a comparable precision when using an automatic method, and possibly a much higher precision when using a semi-automatic method, compared to the conventional TFIDF-based method.

2 First phase: TFIDF-based IR

To separate Japanese words and remove stop words in both queries and documents, original queries and documents were morphologically analyzed by Chasen, i.e., a Japanese morphological analysis tool, and only nouns (including Japanese verbal nouns) were selected for use.

In the first phase, we used two TFIDF-based methods based on Robertson's 2-Poisson model [4].

In the first method, the score of each document is calculated using the following equation,

$$Score(d) = \sum_{\substack{\text{term } t\\ \text{in } d}} \left(\frac{tf(d,t)}{\frac{length(d)}{\Delta} + tf(d,t)} \times \log \frac{N}{df(t)} \right), \tag{1}$$

where t indicates a term that appears in a document, $tf(d,t)$ is the frequency of t in a document d, $df(t)$ is the number of documents in which t appears, N is the total number of documents, $length(d)$ is the length of a document d, and Δ is the average length of the documents.

In the second method, the score of each document is calculated using the following equation, which was first developed by Murata et al. [3, 5].

$$\begin{aligned} Score(d,q) = \sum_{\substack{\text{term } t\\ \text{in } q}} \Bigg(& TF(d,t) \times IDF(t) \times TF_q(q,t) \\ & \times K_{location}(d,t) \times K_{detail} \times \left(\frac{N_q}{qf(t)} \right)^{k_{Nq}} \Bigg) \\ & + \frac{length(d)}{length(d) + \delta}, \end{aligned} \tag{2}$$

where

$$TF(d,t) = \frac{tf(d,t)}{tf(d,t) + k_t \frac{length(d)}{\Delta}}, \tag{3}$$

$$IDF(t) = \log \frac{N}{df(t)}, \tag{4}$$

and

$$TF_q(q,t) = \frac{tf_q(q,t)}{tf_q(q,t) + k_q}. \tag{5}$$

Here, $tf_q(q,t)$ is the frequency of t in a query q, Nq is the total number of queries, and $qf(t)$ is the number of queries in which t occurs. k_t and k_q are constants that are set experimentally. $K_{location}$ and K_{detail} are extended numerical terms that are introduced to improve the precision of results. $K_{location}$ uses the location of the term within the document. If the term is in the title or at the beginning of the body of the document, it is given a higher weighting. k_{Nq} is a parameter, which is set to 1 or 0 for cases of using or not using QIDF (for details see [3, 5]).

3 Second phase: self-organizing documentary maps and ranking relevant documents

In the second phase, the self-organizing map (SOM) proposed by Kohonen [2] was used. 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.

This phase first creates self-organizing documentary maps, in each of which a given query and its relevant documents obtained in the first phase are mapped in order of similarity, i.e., a query 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. This phase then ranks the documents related to the query from the map by calculating the Euclidean distances between the point of the query and the points of the documents in the map and choosing the N closest documents as the final retrieval results.

Because a SOM can only deal with vectors of real numbers, both the query and the documents have to be coded into vectors first. Suppose we have a query Q and a set of relevant documents:

$$A = \{A_i \quad (i = 1, \cdots, a)\}, \tag{6}$$

where a is the total number of documents related to Q. 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 or a document A_i. That is, we define a new set

$$D = \{D_i \quad (i = 1, \cdots, d)\} = Q \bigcup A \tag{7}$$

which includes all queries and documents. Here, d is the total number of queries and documents (i.e., $d = 1 + a$). Each document, D_i, can then be defined by the set of nouns it contains as

$$D_i = \{noun_1^{(i)}, w_1^{(i)}, \cdots, noun_{n_i}^{(i)}, w_{n_i}^{(i)}\}, \tag{8}$$

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,

$$w_j^{(i)} = \mathrm{tf}_j^{(i)} \quad \text{or} \quad \mathrm{tf}_j^{(i)}\mathrm{idf}_j. \tag{9}$$

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

$$w_1^{(i)} + \cdots + w_{n_i}^{(i)} = 1. \tag{10}$$

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

$$V(D_i) = [d_{i1}, d_{i2}, \cdots, d_{id}]^T. \tag{11}$$

The $V(D_i) \in \Re^d$ is the input to the SOM. Therefore, the method for computing 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 the 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.

One way to calculate d_{ij} is as follows:

$$d_{ij} = \begin{cases} 1 - \frac{|C_{ij}|}{|D_i|+|D_j|-|C_{ij}|} & \text{if } i \neq j \\ 0, & \text{if i=j} \end{cases} \tag{12}$$

where $|D_i|$ and $|D_j|$ are values (the numbers of elements) of sets of documents D_i and D_j defined by Eq. (8) 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. (8) for D_i and D_j as follows.

$$\begin{aligned} D_i = \{&(c_1, w_{c1}^{(i)}, \cdots, c_l, w_{cl}^{(i)}), \\ &(n_1^{(i)}, w_1^{(i)}, \cdots, n_{m_i}^{(i)}, w_{m_i}^{(i)})\}, \end{aligned} \tag{13}$$

and

$$\begin{aligned} D_j = \{&(c_1, w_{c1}^{(j)}, \cdots, c_l, w_{cl}^{(j)}), \\ &(n_1^{(j)}, w_1^{(j)}, \cdots, n_{m_j}^{(j)}, w_{m_j}^{(j)})\}, \end{aligned} \tag{14}$$

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. (8) and Eqs. (13) and (14), we know that $l + m_i + m_j = n_i + n_j$. Thus, $|D_i|$ (or $|D_j|$) of Eq. (12) can be calculated as follows.

$$|D_i| = \sum_{k=1}^{l} w_{ck}^{(i)} + \sum_{k=1}^{m_i} w_k^{(i)}. \tag{15}$$

Then, we devised a method for calculating $|C_{ij}|$ as follows.

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

Note that we need not consider the case where both are queries for calculating $|C_{ij}|$ because it does not exist.

4 Experimental Results

4.1 Systems

We submitted the following five systems for the JJ task.

System1:

Eq. (1) was used as the retrieval model, and the document titles and contents were used as retrieval objects. For queries, only titles were used.

System2:

Eq. (1) was used as the retrieval model, and the document titles and contents were used as retrieval objects. For queries, only descriptions were used.

System3:

Eq. (1) was used as the retrieval model, and the document titles and contents were used as retrieval objects. For queries, descriptions and relevant documents (narrative) were used.

System4:

A two-phase system (i.e., Eq. (1)+SOM) was used as the retrieval model, and the document titles and contents were used as retrieval objects. For queries, descriptions and relevant documents (narrative) were used.

System5:

A two-phase system (i.e., Eq. (2)+SOM) was used as the retrieval model, and the document titles and contents were used as retrieval objects. For queries, titles, descriptions, relevant documents (narrative), and concept fields of queries were used.

4.2 Parameters

For the details of the parameters used in Eq. (2), see [6]. The weights (Eq. (9)) were computed using tfidf values. The parameters used in the SOM are as follows. SOMs of 60×60 two-dimensional arrays were used. In the ordering phase, the number of learning

Table 1. Experimental results

Systems	Query	Average precision Relax	Average precision Rigid	R-precision Relax	R-precision Rigid
System1	T	0.2635	0.1802	0.2954	0.1944
System2	D	0.2635	0.1802	0.2954	0.1944
System3	DN	0.3801	0.2933	0.4049	0.3085
System4	DN	0,1625	0.1164	0.1740	0.1194
System5	TDNC	0.1757	0.1193	0.1738	0.1155

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 $\mathbf{m_i(0)}$ consisted of random values between 0 and 1.0.

4.3 Results and Discussions

The experimental results[1], in which using the two-phase system was generally worse than using TFIDF-based methods, are shown in Table 1.

When we used the proposed method to IREX dry run correct data, we obtained good results [1], whereas we obtained very bad results in this NTCIR-5 task. We think that these contrary sets of results were caused by differences between the two experiments. First, the number of documents used in the second phase was different: 439 in the IREX experiment and 1000 in this task. Second, the number of maps to be created and the number of queries mapped in one map were different: only one map (one map for the all six queries and their relevant documents) in the IREX experiment and 47 maps (one map for one query) in this task. Third, the ratio of relevant documents in retrieved documents was different: 0.17 in the IREX experiment and 0.09 in this task on average. Particularly, we also see the case of ratio of 0.011 in this task.

Taking these remarkable differences between the IREX experiment and the NTCIR5 task, to optimally adapt our system to practical IR tasks, we propose the following three solutions. First, we should restrict the number of documents for mapping in the second phase (e.g., instead of the total number of documents that have been retrieved in the first phase, only the top 10% or less of the documents retrieved in the first phase should be used for mapping and re-ranking in the second phase). Second, by considering the differences between queries and documents, we think mapping both the query and documents in a same map is a bad policy. Instead, we should map only documents and rank these documents on the map by calculating the distance between the point of each document and the central point, which may be that of the top relevant document or that of the average of the top N documents determined in the first phase. Third, the number of documents to be treated in the second phase, the number of top documents to be used as the central point in distance calculation, and the size of the SOM should be optimally determined by preliminary experiments using some practical data, which may be the data used in the previous NTCIR tasks.

Because of limited time, we only performed two very beginning additional experiments. The first experiment was based on the first solution, that is, the number of documents for mapping was restricted to the top 100 documents, other than the size of the SOM, which was reduced to 20 × 20, no other parameters were modified, and each query was mapped with the documents. The second experiment was based on the first and second solutions, that is, the number of documents for mapping was restricted to the top 100 documents, other than the size of the SOM, which was reduced to 20 × 20, no other parameters were modified, and no queries were mapped with the documents (instead, the top document retrieved in the first phase was used as the central point in distance calculation). The experimental results showed that the average precision of System4 was improved to 0.2797 (from 0.1625) by using the first solution and further improved to 0.3673 (in Relax case) by using both the two solution. These results showed that the first and second solutions we proposed are greatly effective and our proposed method would have a comparative performance with conventional TFIDF-based methods. We believe that our method will reach much higher precision if top N documents instead of only the top one are used as the central point in distance calculation and the third solution is also used.

By examing the details of the results, we found that for some queries the retrieved results were much better than when using System3, the best one of the systems we submitted (Table 1). For example, there was a case where the precision of the top 10 was 1.0. For other queries, however, the results were much worse than when using System3. For example, the precision of the top 10 was 0.0. We think this means the top 10 documents were mapped in an area far from the query or the top document (which was used as the central point in distance calculation). This gave us two hints. First, there is more work in obtaining an optimal cen-

[1] The results of System1 and System2 were same, which may imply there were some mistakes in our experiments. We will check this later.

tral point for distance calculation. Second, if a semi-automatic IR way is used, that is, if the self-organized map is used to pick up documents that are clustered together in the map as the highly relevant documents, then good results can be obtained even from a map with extremely low precision. In other words, if a semi-automatic method is used, then our two-phase system should have much better performance than the conventional TFIDF methods.

5 Conclusion

We participated in the Single Language IR (SLIR) subtask of CLIR task at NTCIR-5 to verify the practical effectiveness of a two-phase IR system, aimed at visualizing information retrieval while at the same time improving its precision. Although the proposed system had pretty good performance in relatively small computer experiments, the results of this task showed that additional work in adapting our system to large-scale IR tasks is necessary. This paper described the details of our systems submitted to the task and their experimental results, and proposed three solutions for solving the scaling problem so that the proposed system can be effectively adapted to practical IR tasks. Very beginning additional experiments incorporating the solutions showed that our system has a comparable performance when using an automatic method, and has much higher precision when using a semi-automatic method, compared to the conventional TFIDF-based method.

References

[1] Q. Ma, K. Enomoto, M. Murata, and H. Isahara. Information retrieval capable of visualization and high precision, IJCNLP-05, Jeju Island, Korea, Oct., 2005.

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

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

[4] S. E. Robertson and S. Walker. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. ACM SIGIR'94, 1994.

[5] M. Murata, Q. Ma, and H. Isahara. Applying multiple characteristics and techniques in the NICT information retrieval system, The 4th NTCIR Workshop Meeting, Tokyo, Jun., 2004.

[6] M. Murata, Q. Ma, and H. Isahara. Applying multiple characteristics and techniques in the NICT information retrieval system in NTCIR-5, to appear in the 5th NTCIR Workshop Meeting, Tokyo, Dec., 2005.