A Relevant Score Normalization Method Using Shannon's Information Measure*

Yu Suzuki[1], Kenji Hatano[2], Masatoshi Yoshikawa[3],
Shunsuke Uemura[2], and Kyoji Kawagoe[1]

[1] Ritsumeikan University, 1-1-1 Noji-Higashi, Kusatsu, Shiga 525-8577, Japan
[2] Nara Institute of Science and Technology,
8916-5 Takayama, Ikoma, Nara 630-0192, Japan
[3] Nagoya University, Furo, Chikusa, Nagoya, Aichi 464-8601, Japan

Abstract. Given the ranked lists of images with relevance scores returned by multiple image retrieval subsystems in response to a given query, the problem of combined retrieval system is how to combine these lists equivalently. In this paper, we propose a novel relevance score normalization method based on Shannon's information measure. Generally, the number of relevant images is exceedingly smaller than that of the entire retrieval targets. Therefore, we suppose that if the subsystems can clearly identify which retrieval targets are relevant, the subsystems should calculate high relevance scores to a few retrieval targets. In short, we can calculate the sureness of the IR subsystem using the distribution of the relevance scores. Then, we calculate the sureness of the IR subsystems using Shannon's information measure, and calculate the normalized relevance scores using the sureness of the IR subsystems and the raw relevant scores. In our experiment, our normalization method outperformed the others.

1 Introduction

In Web metasearch engine research field, researchers have been discussed how to deal with multiple retrieval results, whereas the researchers in image retrieval research field have discussed rarely. For example, Montague et al. [1, 2] premise an Web IR system that combines multiple retrieval results of some Web IR subsystems. When that Web IR system combines relevance scores, the similarity values between the retrieval targets and the users' queries, the IR system does not combines the raw relevance scores directly, but combines the normalized relevance scores. Because, the relevance scores calculated by different IR subsystems are not always equivalent with each other.

We should note that the high relevance scores calculated by subsystems do not always indicate high relevances. Because, if we have a poor subsystem, this subsystem cannot identify which retrieval targets are relevant to the users' queries. In this case, the subsystem may calculates high relevance scores to many retrieval targets even if these retrieval targets are irrelevant. However, these high relevance scores actually do not indicate the high relevances.

* This work is partially supported by the Ministry of Education, Culture, Sports, Science and Technology, Japan under grants 16700103.

E.A. Fox et al. (Eds.): ICADL 2005, LNCS 3815, pp. 311–316, 2005.

In this paper, we propose a novel relevance score normalization method using the sureness of the IR subsystem. We assume that the sureness depends on the distribution of the relevance scores calculated by the IR subsystem. For example, when the IR system calculates a few high relevance scores, the IR system has a high sureness. From this assumption, we measure the sureness of the IR systems using Shannon's information measure [3]. And then, the IR system combines this measure and the raw relevance scores.

2 Basic Issues

In this section, we introduce the following two issues; 1) why we decide that the IR system combines relevance scores instead of combines feature vectors, and 2) why we decide that the IR system uses the relevance scores instead of ranks.

First, we introduce two typical types of approaches about the IR systems that can deal with multiple features. One approach is a method of combining multiple feature values, which is mainly used in the content-based image retrieval research field [4]. In this method, the IR systems merge multiple feature values into one feature vector per one retrieval target. Therefore, using this combined feature vectors, the IR systems can calculate similarity values between the users' queries and the retrieval targets. Nevertheless, the IR systems do not always deal with the feature values equally. For instance, we suppose that an IR system extracts two kinds of feature vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ from each retrieval target, and the numbers of dimensions of $\boldsymbol{a}$ and $\boldsymbol{b}$ are 1 and 10000, respectively. Of course, the number of dimensions in the merged feature vector $\boldsymbol{c}$ is 10001. In this case, the elements of two vectors, such as $\boldsymbol{b}$ and $\boldsymbol{c}$, are almost the same. Accordingly, the IR systems do not deal with $\boldsymbol{a}$ when the IR systems merge these two feature vectors. In this way, using a merging approach, the IR systems may ignore some feature values.

Here, we are interested in how to equivalently deal with the multiple feature values. We suppose that if the IR systems use better approach, a method of merging multiple retrieval results, the method will be able to improve the accuracies of the IR systems. Before explaining this approach, we show the overview of the image IR system using this method in Figure 1. Our image IR system retrieves the retrieval targets using the following three steps; (1) Our IR system inputs the users' queries to these three IR subsystems, (2) each IR subsystem outputs the retrieval result using one kind of feature value, and (3) the IR system integrates these three retrieval results and outputs one integrated retrieval result. In this case, we suppose that if the IR system equally integrates the retrieval results, the IR system can equivalently deal with the multiple feature values. In this paper, we suppose that step (3) of the IR system is the most important step, then we consider how to integrate multiple retrieval results equivalently.

As mentioned earlier, the goal of this research is to find a retrieval method that can deal with multiple feature values equivalently. To this end, we use the relevance scores instead of the ranks, because the ranks do not always express the exact similarities between the retrieval targets and the users' queries. This means that when we have the two ranks, that are calculated by different IR subsystems, and that are calculated for the two different retrieval targets, are the same, we suppose that the similarities between the two retrieval targets are different in many cases. The reason of this is that the rank

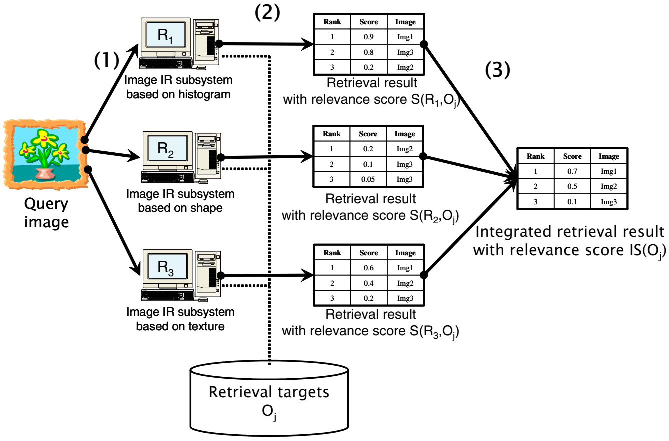
(1)
(2)
(3)
R_1
Image IR subsystem based on histogram
R_2
Image IR subsystem based on shape
R_3
Image IR subsystem based on texture
Query image
Rank Score Image
1 0.9 Img1
2 0.8 Img3
3 0.2 Img2
Retrieval result with relevance score $S(R_1,O_j)$
Rank Score Image
1 0.2 Img2
2 0.1 Img3
3 0.05 Img3
Retrieval result with relevance score $S(R_2,O_j)$
Rank Score Image
1 0.6 Img1
2 0.4 Img2
3 0.2 Img3
Retrieval result with relevance score $S(R_3,O_j)$
Rank Score Image
1 0.7 Img1
2 0.5 Img2
3 0.1 Img3
Integrated retrieval result with relevance score $IS(O_j)$
Retrieval targets O_j

Fig. 1. An architecture of the IR system that uses multiple feature values

of a retrieval target depends on not only the exact similarity value between the retrieval target and the user's query but also the similarity values of the other retrieval targets. Because of the above discussion, we use the relevance scores to calculate the integrated retrieval results.

3 Relevance Score Normalization Method

In this section, we explain a requirement of normalized relevance scores. We also explain our proposed method that fulfill such the requirement.

The basic concept of our idea is that the sureness of the IR subsystems should depend on the distributions of the relevance scores. This means that when an IR subsystem calculates high relevance scores to many retrieval targets, these relevance scores do not necessarily to indicate high relevances, then the IR subsystem should have low sureness. In this section, we explain our normalization concepts in detail, and we also explain how to normalize the relevance scores using the sureness of the IR subsystems.

A Requirement for Relevance Score Normalization Method. Before describe our proposed normalization method, we should discuss which method is the best. To determine the effectiveness of normalization methods, we define that if two normalized relevance scores calculated by any two retrieval targets are the same, users judge that the relevance of one retrieval target is as much as that of another retrieval target.

For example, if two relevance scores, $S(R_1, O_1)$ and $S(R_2, O_2)$, are the same value, users judge whether both two retrieval targets and are relevant or irrelevant to the query. Therefore, when a user judge that O_1 is relevant and O_2 is irrelevant when the IR system uses a normalization method, this method is the best.

When a normalization method fulfill this requirement, the normalization method must determine which two raw relevance scores should be normalized to the same value. Therefore, we suppose that the relevance feedback method is the most suitable method for normalization. This is because, unless the users' judgements, the IR system cannot

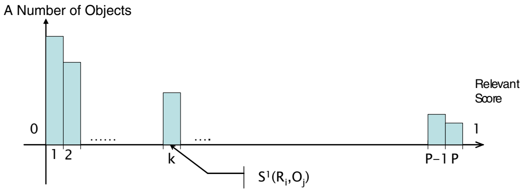

Fig. 2. Calculation method of the information values

identify which retrieval targets are really relevant to the users. However, users cannot judge numbers of retrieval targets. Therefore, we do not adopt the relevance feedback method to normalize raw relevance scores.

To fulfill the above requirement, we need a normalization method that fulfill this requirement without users' judgements. In the following sections, we discuss several normalization methods from the point of view of whether it meets the above requirement.

The idea behind our method. To make our method that fulfill the requirement for normalization, we focus on the distribution of the raw relevance scores. That is, if an IR subsystem calculates many high relevance scores, this IR subsystem has a low sureness. We use Shannon's information measure to scale the difficulty of getting high relevance scores. That is, if a retrieval system calculates many high relevance scores, the information value is low, and the normalized relevance scores are also low.

Calculation of the Information Value of Relevance Scores. Using the pre-normalized relevance scores, the IR system calculates the information values of all retrieval targets. First, we divide the ranges [0, 1] into the p-th sections shown in Figure 2, where p is an integer parameter. We should note that the length of each section L is $1/p$. Next, the IR system set the values of $F(k)$ which expresses a distribution of the relevance scores. Here, the IR system sets the value of $F(k)$, the ranges $[\frac{k}{p}, \frac{k+1}{p}]$, to the number of the retrieval targets which relevance scores are on the range $[0, \frac{k+1}{p}]$. Finally, the IR system calculates the information value $I(R_i, O_j)$ using the following function:

$$I(R_i, O_j) = -\log_2 \frac{F(k)}{M} \tag{1}$$

This function is based on Shannon's information measure [3]. Here, Shannon's information measure is based on the probability of the phenomenon. Therefore, we cannot use this measure directly. Then, we use the ratio of the number of retrieval targets in k-th section to the amount of all retrieval targets instead of the probability of the phenomenon.

Integration of Relevance Scores and the Information Value of Relevance Scores. Finally, the IR system calculates the normalized relevance score using the raw relevance score $S^*(R_i, O_j)$ and the information value $I(R_i, O_j)$ as follows:

$$S'(R_i, O_j) = S^*(R_i, O_j) \cdot I(R_i, O_j) \tag{2}$$

Using these steps, the IR systems can normalize relevance scores using Shannon's information measure.

4 Experimental Evaluation

We compare the accuracy of the IR systems which use our proposed method with that which uses the other normalization methods. We made image retrieval systems which deals with three image features, such as color histogram, shape of objects in the image, and texture of objects.

In our experiment, we compared our proposed normalization methods with the other normalization methods proposed by Montague et al. [1], such as Standard, Sum, and ZMUV. After we normalize, we used two integration functions, CombSUM and CombMNZ, to integrate relevance scores of histogram, shape, and texture. In short, we compared 11 patterns of retrieval systems. Three patterns of systems use one of three image features. Eight patterns of systems use one of four normalization methods, such as Standard, Sum, ZMUV, and our proposed methods. These eight patterns of systems also use one of two integration functions, such as CombSUM and CombMNZ.

In Fig. 3, we show the average precision ratio of all IR systems. From this figure, we find out that the accuracy of the IR system which use our proposed method with CombMNZ gives better accuracy than the other normalization system, such as Standard, Sum, and ZMUV. However, the IR system which uses our proposed method with CombSUM has worth accuracy than that which uses Standard with CombSUM. From this result, our proposed method does not always gives the best accuracy. We suppose that when the IR system uses the integration function CombSUM, the IR system ignores the information measure.

From this result, we can conclude that the compatibility of integration function and normalization method is important. This is because, the method of CombMNZ is based on the entropy of integrated relevance scores, which are very similar to our proposed method. On the contrary, CombSUM is based on average value of integrated relevance scores. This reason is why our proposed method makes better accuracy with CombMNZ

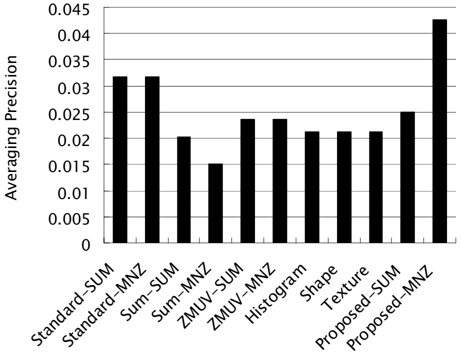

Fig. 3. The 11pt averaging precision of retrieval results using the normalization methods

than with CombSUM, and why the normalization method “SUM” makes better accuracy with CombSUM than with CombMNZ.

5 Conclusion

In this paper, in order to improve the accuracy of the IR system, we introduce a relevance score normalization method. In our method, we expect the sureness of the IR subsystems from the distribution of the relevance scores of the IR subsystems. That is, when the IR subsystems calculates high relevance scores to many retrieval targets, we suppose that these retrieval targets are not always relevant. Based on the sureness of IR subsystems and the raw relevance scores, we calculate the normalized relevance scores. Using our proposed normalization method, the accuracy of the IR system improves without complicated interactions between the IR system and the users. We suppose the reason of the improvement is that we can correctly assume the sureness of the IR system using the distribution of the raw relevance scores and Shannon’s information measure.

References

1. Montague, M., Aslam, J.: Relevance Score Normalization for Metasearch. In: Proceedings of the 10th ACM International Conference on Information and Knowledge Management (CIKM01). (2001) 427 – 433
2. Montague, M., Aslam, J.: Conduct fusion for improved retrieval. In: Proceedings of the 11th ACM International Conference on Information and Knowledge Management (CIKM02). (2002) 538 – 548
3. Shannon, C.E.: A mathematical theory of communication. Bell System Technical Journal **27** (1948) 379 – 423
4. Veltkamp, R.C., Tanase, M., Sent, D.: 5. In: State-of-the-art in Content-Based Image and Video Retrieval. Kluwer Academic Publishers (2001) 97 – 124