RELATIVE RELEVANCE FEEDBACK IN IMAGE RETRIEVAL

Yuki Sugiyama, Makoto P. Kato, Hiroaki Ohshima, Katsumi Tanaka
Department of Social Informatics
Graduate School of Informatics, Kyoto University
Kyoto, Japan
{sugiyama, kato, ohshima, tanaka}@dl.kuis.kyoto-u.ac.jp

ABSTRACT
We propose a relative relevance feedback method for image retrieval systems. Relevance feedback is an effective method to modify a user’s query by selecting relevant and irrelevant items in the search result. However, users cannot always find exactly relevant items in the first few search result pages, especially when the initial query is not specified due to the lack of user’s knowledge. Thus, we propose relative relevance feedback in the present paper, which allows users to select relatively relevant and irrelevant items, and modifies a query by taking into account the relativity of user’s feedback. Our experimental result shows that the relative relevance feedback outperforms a conventional relevance feedback for image retrieval tasks.

Index Terms—relative relevance feedback, image retrieval

1. INTRODUCTION
With the growth of the Internet, various kinds of data have been uploaded on the Web in recent years. Users search for information on a daily basis using a Web search engine. However, it is difficult to input an appropriate query when the user has little knowledge on a domain where he wants to find information. For example, a user who wants to know the name of a bird he saw around a lake might input an underspecified query “water bird” to a Web image search engine, because he does not know the name of that bird. He would feel difficult to find a photo of that bird from millions of search result items returned in response to his initial query. Therefore, many interactive methods have been proposed to help the user modify his query so that he can reach his desired information.

Relevance feedback (RF) [11] is a typical method to support user’s query modification, where users can give feedback to a search system by selecting relevant and irrelevant search result items. A RF system automatically modifies the initial query based on user’s feedback, and presents new search results to the user. However, users cannot always find exactly relevant items in the first few search result pages, especially when the initial query is not specified due to the lack of user’s knowledge. (We shall show how likely this case happens by using a Flickr dataset.) Thus, users may select relatively relevant items as positive examples for the RF, even though those items are absolutely irrelevant. The original RF, which assumes users indicate exactly relevant and irrelevant items, would not work well in this case. Suppose that in the example of the query “water bird” the user is finding a photo of (say) egrets. He might select photos of swans as relatively relevant examples when no egret photo can be found in the first few search result pages. An original RF system, which tries to find images similar to ones selected as relevant, would rank swan photos at the top, which might occupy the first few search result pages in the age of information explosion. Whereas, few egret photos are likely to be included in the top search results, since the similarity between swan and egret photos is not as high as that across swan photos.

In this paper, we propose relative relevance feedback for image retrieval that aims to capture the relativity of positive and negative examples in search result pages. Given positive examples, the only thing we know is that selected items are more similar to an ideal query than items that are not selected as relevant. However, we have no information on exactly where an ideal query point should be set. Thus, our method modifies the initial query point so that a query is move to the centroid of all the items that are more similar to selected items than non-selected ones. This query modification can be explained by using the Voronoi diagram: a new query is located at the centroid of Voronoi cells to which selected items belong when presented search result items are used as seeds for Voronoi diagram generation. We also propose a method to locate a new query point in between selected items and the centroid of Voronoi cells.

The effectiveness of our proposed method was demonstrated through relevance feedback simulations with a Flickr image dataset. We conducted RF simulations on several types of initial search result. The experimental result shows that the relative RF outperforms a conventional RF especially when the initial search result contains only images that are similar to desired images but not relevant.

The rest of this paper is organized as follows. In Section 2, we review related work. Section 3 introduces the relative
relevance feedback model and Section 4 gives more details with an application. We show the experimental results in Section 5 and conclude in Section 6.

2. RELATED WORK

Rocchio proposed a relevance feedback method, which modifies a query by adding positive examples and subtracting negative examples in the vector space model. Rocchio’s RF model has been widely employed in several studies on relevance feedback, and we shall refer to his method as the conventional RF hereinafter. The Multimedia Analysis and Retrieval System (MARS) [12, 13, 14] is an image search system developed by Rui et al. and has a RF system. This system changes a distance metric by calculating the variance of each feature dimension of selected positive examples. Ashwin et al. [1] indicated that if only positive examples are used for distance metric prediction, such as with MARS, items close to non-relevant ones continue to be retrieved. They proposed a robust technique for incorporating non-relevant items to determine the feasible search region. A decision surface is determined to split the attribute space into relevant and non-relevant regions. Many of the RF systems have demonstrated the effectiveness of query modification based on user’s feedback, assuming that relevant search result items are always available within the first few search result pages. In the present paper, we show that it is not likely that relevant items are not always available especially when the initial query is underspecified, and propose a new query modification method that suits for relative RF.

Some papers show retrieval performance can be improved by pseudo-relevance feedback [3, 10], which is also used in multimedia retrieval [4, 17]. This method considers documents ranked at top k as relevant and then applies a traditional RF method. In a situation that the user cannot input an appropriate query, the top results are not always relevant to her ideal search intent, and relevant search result items may be scattered across much lower ranks. In this paper, we try to find candidates for the desired information from users’ behavior.

Joachims et al. [5, 6] focused on unclicked links in clickthrough data and used them to improve the retrieval quality of search engines. The strategies they explored are based on the idea that a clicked link is more relevant than skipped links ahead of itself. We consider the links that a user viewed in first few result pages as representatives of whole result links and use the links represented by the clicked one to modify the query.

Nakajima and Tanaka proposed a relative query for image search [8], by which users can select relatively relevant images as a query from an image collection. They calculate the similarity between input queries and data to be retrieved taking into account the relativity of the selected images in the image collection. Although their work is similar to ours, it does not focus on relevance feedback.

Fig. 1. Prospective user’s action. When the user can view only high-ranked objects, we assume that he is most likely to select the object closest to a desired object.

Fig. 2. Voronoi diagram. A feature space is divided by high-ranked objects. Objects in the same Voronoi cell have the potential of user’s desired information.

3. RELATIVE RELEVANCE FEEDBACK

3.1. Definition of Retrieved data and Query

In our model, data to be retrieved such as images and Web pages are called objects, which are denoted by \( O = \{o_1, o_2, \ldots, o_n\} \). Each object \( o \) is represented as a feature vector, which we describe as \( f(o) \). A query is also represented as a feature vector \( q \), and a query at time \( t \) is denoted by \( q_t \). For example, \( q_0 \) represents the initial query and \( q_2 \) represents a modified query after second feedback. Given a query \( q_t \), objects \( O \) are ranked based on the similarity between the query and their vector.

3.2. Relative Selection

When a user starts search, initial search result are shown anyway. The relevance feedback allows him to give a feedback on the result. In many case, the initial result does not have exactly relevant items but has a relatively relevant item. The user has no choice but to select relatively the most relevant item. In this case, traditional RF techniques based on Rocchio’s formula would retrieve objects exactly similar to the selected items. The user’s desired objects are not necessarily close to the selected item but relatively closer to the selected item. In the other word, we can assume that the
selected item is closer to the user’s desired objects than the other unselected items (Fig. 1).

On this assumption, we borrow an idea from the Voronoi diagram (Fig. 2). It is a decomposition of a space in which the distance between two objects is defined. When \( k \) seed points are given, the space is divided into \( k \) regions based on the idea that which seed is the closest to the position. When we regard objects user see in the search result as seed points, objects in the Voronoi cell generated by the selected one are candidates of the user’s desired objects.

### 3.3. Query Modification

When a set of sample objects \( O_S \) is shown as search result, the user can judge whether each of them is relevant to his search intent or not. As previously mentioned, the judgment may be relative. We denote a set of the relatively positive examples by \( O_P \), and a set of negative ones by \( O_N \).

\[
\begin{align*}
O_P &= \{ o \in O \mid o \text{ is shown to a user} \} \quad (1) \\
O_S &= \{ o \in O_S \mid o \text{ is relatively positive} \} \quad (2) \\
O_N &= \{ o \in O \mid o \text{ is relatively negative} \} \quad (3)
\end{align*}
\]

If Rocchio’s original method is applied to the user’s positive and negative feedback, the query modification formula would be as follows.

\[
q_{t+1} = \alpha q_t + \frac{\beta}{|O_P|} \sum_{o \in O_P} f(o) - \frac{\gamma}{|O_N|} \sum_{o \in O_N} f(o) \quad (4)
\]

Here, coefficients \( \alpha, \beta, \gamma \) are the weights of the original query, positive and negative feedback. The modified query by this formula is to obtain items that are exactly similar to the positive examples and obviously far from the negative examples.

Our proposed relative relevance feedback considers not only user’s selected items but items that are not shown to the user. Thus preferred formula is

\[
q_{t+1} = \alpha q_t + \frac{\beta}{|O_P|} \sum_{o \in O_P} g(o, O_S) - \frac{\gamma}{|O_N|} \sum_{o \in O_N} g(o, O_S) \quad (5)
\]

where a feature vector \( f(o) \) is transformed into another vector by mapping \( g \). Especially, when \( g \) returns the feature vector of the input object \( o \) without any change, our formula is equivalent to Rocchio’s. In other words, our method expands the traditional relevance feedback.

### 3.4. Mapping algorithm

In order to retrieve not only images exactly similar to selected one but also relatively relevant ones, we have to modify the query in consideration of the images that are more similar to selected images than unselected ones. We can get these images by using the Voronoi diagram discussed in Section 3.2. According to the definition of the diagram, images in the Voronoi cell generated by the user’s selected image just meet the requirement.

One of the most representative points of a certain region or a set of points is the centroid of it. Thus, in the relative relevance feedback, the query vector has to be modified to be closer to this point than the selected point. How we should bring the query close to the centroid depends on how the selected image is absolutely relevant to the ideal search intent. We decide that the appropriate point is situated between the selected images and the centroid of Voronoi cells. Therefore, we define the mapping \( g \) as follows

\[
g(o, O_S) = (1 - \mu)f(o) + \frac{\mu}{|V(o, O_S)|} \sum_{o' \in V(o, O_S)} f(o') \quad (6)
\]

Here, the parameter \( \mu \) is the ratio of internal division, and \( V(o, O_S) \) is a set of images in the Voronoi cell generated by the seed image \( o \). It is defined as follows.

\[
V(o, O_S) = \left\{ o \left| \min_{o' \in O_S} \| f(o), f(o') \| = o \right. \right\} \quad (7)
\]

### 3.5. Ranking of Search Result

There are many methods to show search results to the user. A ranking function that quantifies the relevance between a target object and a query, for example, is often used to define the order of objects shown in search result. In information retrieval, the best order of search result is not determined. An order by relevance between the query vector and feature of each object is sometimes important. And other times, it is better to increase the diversification of the result list.

In this section, therefore, we define that the ranking function \( \text{Rank} \) which determine the order of each object \( o \in O \) is decided by the current query \( q_t \), and describe as follows.

\[
\text{Rank}(o, q_t) \quad (8)
\]

### 4. APPLICATION

We implement a sample image retrieving application which enables the relative relevance feedback with images posted on Flickr\(^1\), an image-hosting Web service.

#### 4.1. Image Feature

Every image needs to be represented as a feature vector in our application. There are many kinds of methods to construct a vector from an image. We utilize the visual vocabulary tree based method proposed by Nister and Stewenius [9]. In this method, first the feature points are extracted from an image. They originally used SIFT [7] for the feature extraction. Instead, we use SURF [2] because it

\(^{1}\text{http://www.flickr.com/}\)
extracts feature points faster than SIFT. Each image is represented as a vector with about one million dimensions.

4.2. Feedback Operation

As previously discussed, a user cannot see all of the search result images. Our application shows 12 images per page, and we regard the first page as the subset the user is able to check. The user judges the relevance of each image on the first page and then re-search relatively relevant to positive examples and relatively irrelevant to negative ones.

Simple negative feedback sometimes impairs the accuracy when the search result has many irrelevant images [16]. Vasconcelos and Lippman noted that negative feedback is useful when the user has already given the system all the positive examples, and the system returns exactly the same images as ones in the previous iteration [15]. Since we assume the situation that the user cannot check all of the relevant images, negative feedback may not work well. Thus, we use only positive feedback for simplification. Relevance feedback modifies the query vector. The modified query is made by mixing the current query vector and the vector indicated by the feedback. The mixture ratio can be variable. We use

\[ \alpha = 0.5, \beta = 0.5, \gamma = 0 \]

4.3. Ranking method

In image retrieval, there are several ranking methods that determine the order of result images. Some simple methods use \( L_p \)-norm, inner product or cosine similarity of the query and each image feature vector. In this application, we use cosine method, which achieved the best accuracy in some trials. Let \( \mathbf{a} \cdot \mathbf{b} \) stands for inner product of vector \( \mathbf{a} \) and \( \mathbf{b} \), and \( \| \mathbf{a} \| \) denotes the Euclidean distance, then we defined the ranking function as follows.

\[ \text{Rank}(o,q_t) = \frac{f(o) \cdot q_t}{\|f(o)\| \|q_t\|} \]

4.4. Execution Example

Figure 3 shows an example of the execution procedure. In this example, a user wants to search images of egrets even though he does not know the name of “egret.” Since he knows that the birds live around water, he inputs the query “water bird” to our system. Of course, images of various kinds of water bird are initially returned and the result is too large for him to check totally. When he finds an image of the birds (e.g. swans) that is a little relevant to egrets among lots of irrelevant images, he tries to feedback it because it is relatively relevant in the images he has ever seen.

In this situation, the traditional relevance feedback cannot help but retrieve images of swans. On the other hand, our method can retrieve egrets, domestic ducks, etc. which are more similar to swans than unselected birds.

5. EXPERIMENT

We demonstrate the effectiveness of our relative relevance feedback method through simulations with a Flickr image dataset.

5.1. Tasks

We devised 14 image retrieval tasks shown in columns initial query and desired image in Table 1. The goal of each task is to search for a specific kind of image or ones in a specific situation. Since we assume a situation that the user cannot describe desired images specifically, the initial query, which is input by the user at the very beginning of a search session, is underspecified or incomplete. Therefore, search results returned in response to the initial query are not likely to contain desired images. Note that the initial queries were derived from several classes: animals, plants, landscapes and artifacts.

5.2. Dataset

We gathered images from Flickr, an image hosting website, and used them as a dataset for this experiment. By using the Flickr API service, we collected as many images for each initial query as possible under a limitation of the API. The number of images is about 4,500 per query. We recruited volunteers for evaluating the relevance to desired images, and asked them to label all the images as from 0 (irrelevant) to 4 (highly relevant). For reducing the subjectivity of relevance judgment, the intermediate three grades were also explained to the volunteers in natural language: 1 (irrelevant but partially similar to desired images in terms of a certain aspect such as color and texture); 2 (irrelevant but similar to desired images); and 3 (partially relevant). Table 1 shows the statistics of our dataset. It can be seen that the proportion of images with a grade 4 is much smaller than ones with a
We continued simulations until the third feedback is given. For the distance metric, we opted to use cosine similarity after several trial experiments.

5.4. Experimental Results

The main objective of our experiment is to answer a question whether and when our relative RF method can outperform a conventional one. To examine the effect of our method, we conducted RF simulations for different values of the parameter $\mu$. Recall that our proposed method reduces to a conventional RF method, the Rocchio’s method, if the parameter $\mu$ equals to 0. Normalized discounted cumulative gain (nDCG) was used for a metric of image retrieval evaluation that considers the graded relevance and the order. As the initial ranking and selection of positive examples were randomized, we report the average of nDCG@12 over ten runs.

Figure 4 shows nDCG after the third feedback for each value of $\mu$ and each initial grade. This result shows that our proposed method ($\mu > 0$) generally obtained higher nDCG than the Rocchio’s method ($\mu = 0$) when the initial top $k$ search result contains only images with a low grade (grade 1 or 2). Although we cannot find an obvious optimum of the parameter $\mu$ in our simulations, it can be seen that our method with $\mu = 0.4$ performed the best for simulations that start with an image with grades 1 or 2. Whereas, our proposed method ($\mu > 0$) could not obtain higher nDCG than the Rocchio’s method ($\mu = 0$) when the initial grade is 3 or 4. As the relative RF method was developed for user’s feedback where selected search result items are relatively relevant but absolutely irrelevant, it was expected that our method with an initial grade 3 or 4 did not work as well as a conventional RF method. The flattening of the centroid vector cause the decrease of nDCG when $\mu = 1$. Since the number of seed images is much fewer than that of all images and dimension of image feature vector, a Voronoi cell often has a lot of diversified images and the centroid vector becomes flattened.

grade 1 or 2. As discussed earlier in this paper, this result shows that users are likely to fail to find exactly relevant images (grade 4) in the first few search results when the initial query is not specified enough. On the other hand, we could expect that users can find at least similar images to desired ones within a small number of search result images.

5.3. Relevance Feedback Simulation

We summarize how we conducted simulations of relevance feedback in our experiment.

1) A system returns top $k$ search result images in response to the initial query of a task. We don’t have any criteria to rank images at the beginning of a simulation, as the input at this stage is only the initial keyword query. Thus, a system returns all the images tagged by the same term as the initial query, and randomly ranked those images. In our experiment, we conducted simulations manually changing the initial search result. Specifically, we simulated a case that the user can find an image with a grade $x$ within top $k$ search result images, by randomly selecting an image with a grade $x$ and randomly selecting $(k - 1)$ irrelevant images (grade 0). The parameter $x$ is referred to as initial grade in our experimental result.

2) We determine positive examples from top $k$ search result images assuming that users select all the images with the highest grade. A positive example is randomly selected when only irrelevant images are included in the top search results.

3) A system modifies a query according to a query modification method discussed in Section 3.3.

4) A system returns a new search result in response to the modified query based on the ranking function described in Section 3.4.

5) Repeat processes from 2 to 4.

<table>
<thead>
<tr>
<th>Initial query</th>
<th>Desired image</th>
<th>Total size</th>
<th>Grade 1</th>
<th>Grade 2</th>
<th>Grade 3</th>
<th>Grade 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>dog</td>
<td>Beagle</td>
<td>4514</td>
<td>66 (1.5%)</td>
<td>18 (0.4%)</td>
<td>12 (0.3%)</td>
<td>20 (0.4%)</td>
</tr>
<tr>
<td>cat</td>
<td>Persian</td>
<td>4545</td>
<td>114 (2.5%)</td>
<td>76 (1.7%)</td>
<td>26 (0.6%)</td>
<td>26 (0.6%)</td>
</tr>
<tr>
<td>water bird</td>
<td>egret</td>
<td>4484</td>
<td>181 (4.0%)</td>
<td>101 (2.3%)</td>
<td>53 (1.2%)</td>
<td>66 (1.5%)</td>
</tr>
<tr>
<td>freshwater fish</td>
<td>goldfish</td>
<td>4492</td>
<td>750 (16.7%)</td>
<td>124 (2.8%)</td>
<td>18 (0.4%)</td>
<td>51 (1.1%)</td>
</tr>
<tr>
<td>butterfly</td>
<td>Cabbage White</td>
<td>4479</td>
<td>107 (2.4%)</td>
<td>31 (0.7%)</td>
<td>19 (0.4%)</td>
<td>44 (1.0%)</td>
</tr>
<tr>
<td>tree flower</td>
<td>plum</td>
<td>4573</td>
<td>524 (11.5%)</td>
<td>390 (8.5%)</td>
<td>19 (0.4%)</td>
<td>14 (0.3%)</td>
</tr>
<tr>
<td>China waterfall</td>
<td>Huangguoshu</td>
<td>4497</td>
<td>1383 (30.8%)</td>
<td>228 (5.1%)</td>
<td>61 (1.4%)</td>
<td>136 (3.0%)</td>
</tr>
<tr>
<td>New Zealand farm</td>
<td>flock of sheep</td>
<td>4491</td>
<td>192 (4.3%)</td>
<td>185 (4.1%)</td>
<td>58 (1.3%)</td>
<td>24 (0.5%)</td>
</tr>
<tr>
<td>Europe church</td>
<td>Church of the Savior on Blood</td>
<td>4518</td>
<td>44 (1.0%)</td>
<td>8 (0.2%)</td>
<td>7 (0.2%)</td>
<td>8 (0.2%)</td>
</tr>
<tr>
<td>tower</td>
<td>Leaning Tower of Pisa</td>
<td>4507</td>
<td>67 (1.5%)</td>
<td>11 (0.2%)</td>
<td>18 (0.4%)</td>
<td>15 (0.3%)</td>
</tr>
<tr>
<td>Gogh art</td>
<td>Café Terrace at Night</td>
<td>4344</td>
<td>87 (2.0%)</td>
<td>5 (0.1%)</td>
<td>4 (0.1%)</td>
<td>16 (0.4%)</td>
</tr>
<tr>
<td>sports car</td>
<td>Ferrari</td>
<td>4501</td>
<td>1000 (22.2%)</td>
<td>272 (6.0%)</td>
<td>250 (5.6%)</td>
<td>345 (7.7%)</td>
</tr>
<tr>
<td>brass instrument</td>
<td>horn</td>
<td>4497</td>
<td>172 (3.8%)</td>
<td>53 (1.2%)</td>
<td>25 (0.6%)</td>
<td>12 (0.3%)</td>
</tr>
<tr>
<td>Japanese food</td>
<td>tempura</td>
<td>4499</td>
<td>39 (0.9%)</td>
<td>9 (0.2%)</td>
<td>36 (0.8%)</td>
<td>26 (0.6%)</td>
</tr>
</tbody>
</table>

We determined positive examples from top $k$ search result images assuming that users select all the images with the highest grade. A positive example is randomly selected when only irrelevant images are included in the top search results.
images that are similar to desired images but not relevant, especially when the initial search result contains only irrelevant items, and modify a query by taking into account the relativity of user’s feedback. The effectiveness of our relative relevance feedback method when positive examples are totally different, since the mean of centroids of Voronoi cells for such positive examples may cancel out each other. Thus, we will take into account features the user focuses on, and decompose a feature space based on his interest.

In future work, we are planning to improve query modification for multiple positive examples. In the present study, we decompose a feature space into a Voronoi cell for each positive example. However, this might lessen the effectiveness of our relative relevance feedback method when positive examples are totally different, since the mean of centroids of Voronoi cells for such positive examples may cancel out each other. Thus, we will take into account features the user focuses on, and decompose a feature space based on his interest.

6. CONCLUSIONS AND FUTURE WORK

In this paper, we proposed a relative relevance feedback method, which allows users to select relatively relevant and irrelevant items, and modify a query by taking into account the relativity of user’s feedback. The effectiveness of our proposed method was demonstrated through RF simulations with a Flickr image dataset. The experimental result showed that the relative RF outperforms a conventional RF especially when the initial search result contains only images that are similar to desired images but not relevant.

In future work, we are planning to improve query modification for multiple positive examples. In the present study, we decompose a feature space into a Voronoi cell for each positive example. However, this might lessen the effectiveness of our relative relevance feedback method when positive examples are totally different, since the mean of centroids of Voronoi cells for such positive examples may cancel out each other. Thus, we will take into account features the user focuses on, and decompose a feature space based on his interest.

ACKNOWLEDGMENT

This work was supported in part by KAKENHI (No.10J04687) and a Kyoto University GCOE Program entitled “Informatics Education and Research for Knowledge-Circulating Society.”

REFERENCES