# User Term Feedback in Interactive Text-based Image Retrieval

Chen Zhang Joyce Y. Chai Rong Jin
Department of Computer Science and Engineering
Michigan State University
East Lansing, MI 48824
{zhangch6, jchai, rongjin}@cse.msu.edu

## ABSTRACT

To alleviate the vocabulary problem, this paper investigates the role of user term feedback in interactive text-based image retrieval. Term feedback refers to the feedback from a user on specific terms regarding their relevance to a target image. Previous studies have indicated the effectiveness of term feedback in interactive text retrieval [14]. However, the term feedback has not shown to be effective in our experiments on text-based image retrieval. Our results indicate that, although term feedback has a positive effect by allowing users to identify more relevant terms, it also has a strong negative effect by providing more opportunities for users to specify irrelevant terms. To understand these different effects and their implications on the potential of term feedback, this paper further presents analysis of important factors that contribute to the utility of term feedback and discusses the outlook of term feedback in interactive text-based image retrieval.

## Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval]:** Query Formulation, Relevance Feedback, Search Process
H.5.2 [**User Interfaces]:** Interaction Styles

## General Terms

Experimentation, Performance, Human Factors

## Keywords

Interactive image retrieval, user term feedback, text-based image retrieval

## 1. INTRODUCTION

Given tremendous amount of image data, capabilities to support efficient and effective image retrieval have become increasingly important. There are two general approaches for image retrieval. The text-based approaches apply traditional text retrieval techniques to image annotations or descriptions.


SIGIR'05, August 15–19, 2005, Salvador, Brazil.

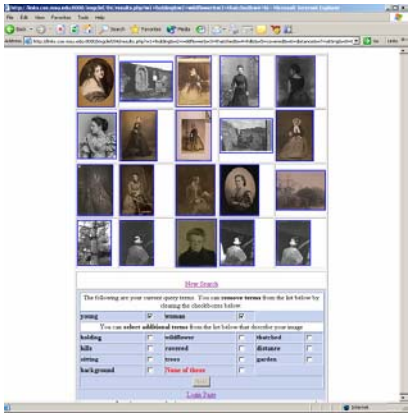

The content-based approaches apply image processing techniques to extract image features and retrieve relevant images. Despite tremendous improvement in content-based retrieval, the content-based approaches still have many limitations. First, it is difficult for users to specify visual queries with low-level visual features. Often, specific tools are required to support queries with rudimentary sketches or paintings [19]. Second, low level image features cannot precisely describe user information needs. There is a gap between low-level visual descriptions and a user's semantic expectation [23]. Text queries, on the other hand, are more intuitive and natural for users to specify their information needs.

However, text-based image retrieval also faces many challenges. One major problem is that the task of describing image content is highly subjective. The perspective of textual descriptions given by an annotator could be different from the perspective of a user. A picture can mean different things to different people. It can also mean different things to the same person at different time. Furthermore, even with the same view, the words used to describe the content could vary from one person to another [13]. In other words, there could be a variety of inconsistencies between user textual queries and image annotations or descriptions.

To alleviate the inconsistency problem, different strategies can be potentially applied, for example, showing textual descriptions of similar images to shape user's query terms; and asking users for relevance feedback concerning the retrieved images. In this paper, rather than investigating these different strategies, we focus on the role of term feedback in interactive image retrieval. Term feedback refers to the feedback from a user on specific terms regarding their relevance to the target information. Previous studies have shown that user term feedback has been effective in interactive text retrieval. In particular, the feedback acquired from users can improve the system final retrieval performance [14]. Therefore our focus in this paper is on whether the use of term feedback can be effective in interactive text-based image retrieval.

Our hypothesis is that users should be more responsive to the terms prompted by the system. It would be easier for users to select terms than to generate terms to describe their target images. Furthermore, the selected terms could bridge the vocabulary discrepancy between image annotations and query terms. Thus the selected terms could potentially improve the

**Figure 1. (a) Manual Refinement Interface**



**(b) Term Feedback Interface**

quality of query expansion, which would ultimately improve the final retrieval results. Based on these hypotheses, we conducted a series of studies. The results from these studies have indicated that term feedback does allow users to identify more terms that are relevant to a target image. However, term feedback also provides more opportunities for users to specify irrelevant terms. Because its negative effect appears to be stronger than its positive effect, term feedback has not shown to be effective in our experiments. To go beyond our current system and understand the implications of these effects on the potential of term feedback, we conducted a further investigation based on both empirical user studies and simulation studies. Our results have indicated that term feedback may not be an effective strategy for interactive text-based image retrieval given the current state-of-the-art. This paper presents our empirical findings, analyses important factors that contribute to the utility of term feedback, and discusses the outlook of user term feedback in interactive text-based image retrieval.

## 2. RELATED WORK

Information retrieval is an interactive process where users play an important role in finding relevant information [10]. Previous studies on interactive text retrieval range from query expansion mechanism [1, 7] to interface design issues [2, 9]. For interactive content-based image retrieval, studies range from visual query construction [19] to relevance feedback on retrieved images [8, 11, 22, 24].

Relevance feedback is an important concept in interactive information retrieval. The idea is that based on the feedback from a user during the retrieval process about the previously retrieved objects, the system can adjust the query to better represent the user's information needs. Previous studies on relevance feedback during interactive text retrieval indicate that users prefer to have explicit knowledge about which terms have been added to the query and also more strongly, users prefer to control which terms to add to the query [3]. The "penetrable relevance feedback" which allows users to add none, some, or all of the suggested terms prior to query expansion has been significantly more effective compared to the systems without such a feature [14]. It has also shown that users like the feedback component since the task of generating terms is replaced by the easier task of term selection [14]. This penetrable relevance feedback has motivated our work in this paper.

For text-based image retrieval (which is the focus of this paper), textual descriptions or annotations have to be provided for images in the first place. To provide text descriptions or annotations, two approaches can be applied. The first approach acquires descriptions/annotations manually by human annotators. The second approach is to automatically annotate images using machine-learning techniques that learn the correlation between image features and textual words from the examples of annotated images [4, 12, 17, 18, 20]. In our work reported here, the text descriptions for images are provided by human annotators.

## 3. EXPERIMENT OF TERM FEEDBACK

We conducted a series of studies to investigate the role of term feedback in interactive image retrieval. The task used in these studies is similar to the task proposed by the CLEF interactive retrieval track [6]: finding a target image in mind through interactive search. In this task, a set of images are put aside as target images. Users are asked to interact with the system to find each of these target images one at a time. This task is somewhat different from the traditional image retrieval tasks. The traditional tasks require users to specify general information needs. In this task, a user already has a target image in mind. Searching for this specific target image usually requires more effort from the user to describe the content of the target image. Therefore, this task is particularly suitable for studying term feedback.

### 3.1 System and Interface

The dataset used in our studies is Eurovision St Andrew collection (provided by the St Andrews University Library) through CLEF (Cross Language Evaluation Forum)[1] [6]. This is a collection of around 30,000 photographs with significant historic value. Each image has a semi-structured caption that

[1] http://ir.shef.ac.uk/imageclef2004/interactive.html

**Table 1: Overall performance of two interfaces**

| | Manual refinement | Term feedback |
|---|---|---|
| Success Rate | 0.48 | 0.27 |
| Average time | 1:41 | 2:41 |
| Average iterations | 4.3 | 4.1 |

consists of information such as the title, textual descriptions of the image content (81% of images have textual descriptions), the date when the photograph was taken, the location of the photograph, etc.

The backend retrieval system applies statistical language model for text-based retrieval [15, 16, 21]. This approach assumes that each description of an image is generated from a random source and its relevance to a given textual query is computed as the likelihood for the underlying random source to generate the query.

Given this backend retrieval system, we implemented two interfaces as shown in Figure 1. Figure 1(a) is a manual refinement interface. Through this interface, users can freely refine their queries (i.e., adding or removing query terms) at each iteration based on the retrieved results. Figure 1(b) is the term feedback interface. In this interface, two lists of terms are shown to the user. The first list of terms captures all the query terms that have been used so far up to this point. The users can de-select any terms from this list after seeing the retrieved results. The second list of terms is automatically generated by the system at each iteration based on user's prior queries and the retrieved results. Details on generating this list of terms will be described in Section 4.2. Users can choose terms from this list to further refine their queries. Having both "de-selecting" list and "selecting list", the term feedback interface is comparable to the manual refinement interface where users can freely add/remove their query terms at each iteration. For both the manual refinement interface and the term feedback interface, a user is required to provide an initial query to start the search. At each iteration, both interfaces show the twenty top retrieved images. The only difference between two interfaces is query refinement as described above.

## 3.2 Methods

Eight subjects participated in the study. Each of them was asked to search for 16 images from the Eurovision St. Andrew collection provided by ImageCLEF. The subjects were first asked to complete a screening questionnaire to elicit demographic data and data concerning searching experience. Then the subject was asked to use one interface to search eight images (one at a time). After using each interface, the subject was given a questionnaire to indicate how easy he/she felt about the search process, and how satisfied he/she was with a particular system. The sequence of which image and which interface to use was predefined based on the Latin Square design in order to guarantee balance between system order and image order. During each search, the system also automatically logged information such as the initial query from a user, the system retrieved results, terms prompted by the system, and the time spent on searching, etc. When an image was found or when five minutes were ran out, the search stopped. After searching all images using two different interfaces, each subject was asked to give an overall ranking of the two interfaces in terms of their overall satisfaction and systems' effectiveness of locating the target images.

## 3.3 Experimental Results

Before conducting the user study, our expectation was that the term feedback interface should work better than the manual refinement interface. This expectation was partly based on our hypotheses and partly based on results from previous studies on term feedback in text retrieval.

However, our experimental results indicate otherwise. Table 1 shows the comparative results for the two interfaces. Since two out of the sixteen images do not have textual descriptions (which are the only information used to generate terms), so the results reported here are based on the fourteen images. The successful retrieval rate is calculated by dividing the number of target images that are shown in the top 20 retrievals by the total number of target images tried for that interface. The success rate for the term feedback interface is significantly lower than the success rate for the manual refinement interface (T-test, $p<0.05$). Among the successful retrievals, the average time spent on retrieving a target image in the term feedback interface is significantly higher than the time spent with the term manual refinement interface. This time difference is potentially due to two factors. The first factor relates to the system computation time required to generate terms in the term feedback interface. The second factor relates to the user response time (i.e., either to manually refine terms or to select terms prompted by the system). In our current investigation, we have not made distinctions between those two contributing factors. Furthermore, as shown in Table 1, among the successful retrievals, the average number of iterations taken to retrieve the target images by the term feedback interface is comparable to that taken by the manual refinement interface.

To further understand these results, we made an effect comparison between two interfaces concerning the *desired* or *undesired* terms given by a user. The *desired terms* are the query terms that appear in the description of target images. The *undesired terms* are the query terms that do not appear in the target descriptions. Table 2 shows the average percentage of *desired* or *undesired* query terms (with respect to the total number of terms in the description of the target image) that are specified among all iterations through each interface.

Our results indicate that the percentage of desired terms is

**Table 2: Effect comparison between the manual refinement interface and the term feedback interface**

| | Manual refinement | Term feedback |
|---|---|---|
| Avg. percentage of desired terms | 0.042 | 0.062 |
| Avg. percentage of undesired terms | 0.161 | 0.183 |

higher in the term feedback interface compared to the manual refinement interface. However, the percentage of undesired terms is also higher in the term feedback interface compared to the manual refinement interface. Furthermore, the percentage of undesired terms is higher than the percentage of desired terms on average. The results imply that in terms of the ability to help users identify relevant (i.e., desired) terms, the term feedback interface works better than the manual refinement interface. However, when terms are provided by the system, there is also a tendency for users to select terms that are irrelevant to the target images. In fact, the undesired terms seem to have a stronger impact on the final retrieval performance than the desired terms. In other words, the negative effect appears to bypass the positive effect for the term feedback interface compared to the manual refinement interface. Therefore, we have not seen the advantage of term-feedback interface.

## 3.4 Factor Analysis

To better understand the positive and negative effects, we analyzed factors that contribute to these effects. We use two variables *PE* (i.e., the positive effect) and *NE* (i.e., the negative effect) to measure the utility of term feedback.

More specifically, for each target image *t*, a given initial query *q*, at each iteration *i*, there will be a *PE* number and a *NE* number. The *PE* indicates the number of terms selected by the user that actually appear in the description of the target image (i.e., the desired terms). The *PE* can be further decomposed as the following three components:

$$PE_{t,q,i} = N_{list} \times GenRate\_des_{t,q,i} \times SelRate\_des_{t,q,i}$$

Here $N_{list}$ is the number of candidate terms shown in the term list. *GenRate_des* is the *desired term generation rate*, which indicates the percentage of terms in the term list that actually appear in the description of the target image. We call these terms *relevant terms*. *SelRate_des* is the *desired term selection rate*, which measures the percentage of the relevant terms that are selected by a user.

Similarly, we can measure the negative effect (*NE*) as follows:

$$NE_{t,q,i} = N_{list} \times GenRate\_undes_{t,q,i} \times SelRate\_undes_{t,q,i}$$

The *NE* indicates the number of terms that are selected by the user but are not in the description of the target image (i.e., undesired terms). Here $N_{list}$ is the same constant as described above. *GenRate_undes* is the *undesired term generation rate*, which measures the percentage of terms in the term list that do not appear in the description of the target image. These terms are called *irrelevant terms*. *SelRate_undes* is the *undesired term selection rate*, which measures the percentage of irrelevant terms that are selected by the user.

It is worth mentioning that the term selection rate in one iteration can affect the term generation rate in the next iteration. To simplify the analysis, we did not take such in-between-iteration effect into consideration. But rather, we incorporated such effect in an average rate across all iterations. Table 3 shows the empirical results of these measurements from our user study. As indicated in Table 3, the product of the average *GenRate_des* and the average *SelRate_des* amounts to 0.016; while the product of the *GenRate_undes* and the average *SelRate_undes* amounts to 0.075. This indicates that the negative effect *NE* is bigger than the positive effect *PE*, which is consistent with our experimental results shown in Table 2.

**Table 3. Empirical data on different effects of term feedback**

| | |
|---|---|
| Average *GenRate_des* | 0.06 |
| Average *GenRate_undes* | 0.94 |
| Average *SelRate_des* | 0.26 |
| Average *SelRate_undes* | 0.08 |

Given ineffective term feedback in our system, the next question is whether this result is largely due to our current design of the system? In other words, is it because of the limitations of our system, the effect of term feedback cannot be clearly demonstrated? If we made certain components better in our system, will the term feedback become more effective? With these questions in mind, we conducted further investigation on the potential of term feedback.

# 4. POTENTIAL OF TERM FEEDBACK

The above discussion has indicated that three factors - the number of terms shown to the user ($N_{list}$), the term generation rate (i.e., *GenRate_des* and *GenRate_undes*), and the term selection rate (i.e., *SelRate_des* and *SelRate_undes*) contribute to the utility of the term feedback interface. Since $N_{list}$ can be pre-defined empirically, so here we specifically focus on the other two factors. Corresponding to these two factors, two directions can be pursued in order to make term feedback effective. The first direction focuses on improving the term selection rate, namely, increasing the chances of selecting the relevant terms from the term list (i.e., *SelRate_des*) and decreasing the chances of selecting the irrelevant terms (i.e., *SelRate_undes*). The second direction focuses on improving the term generation rate, namely, increasing the relevant terms (i.e., *GenRate_des*) and decreasing irrelevant terms in the term list (i.e., *GenRate_des*). To go beyond our current system and understand the potential of term feedback, it is important to investigate the potential of these directions.

## 4.1 Term Selection

The first direction is to identify good strategies to improve the term selection rate (*SelRate_des* and *SelRate_undes*). Our question is whether users respond to different terms differently? Some terms may be more receptive than others. For example, the term "house" could be easier to recognize than the term "background". It seems relatively easier to tell whether there is any house in a target image than to tell what is "background". Can we classify a set of terms that are more "informative" than the others in terms of the ease of recognition with respect to an image? If we can identify the behavior of those "informative" terms, then we can design

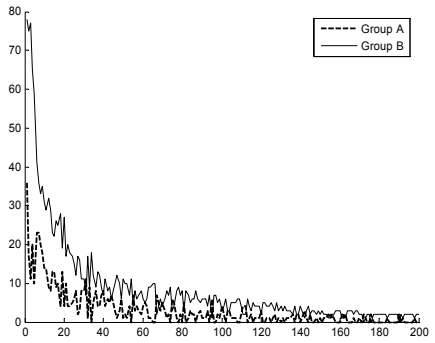
some mechanisms to give those terms higher priority to be included in the term list. Because supposedly users are more responsive to the informative terms, our hypothesis is that such classification can potentially increase *SelRate_des* and decrease *SelRate_undes)*.

Based on these ideas, we conducted a second user study to examine the potential classification of terms for the term selection purpose. Totally 130 users participated in this study through the web. These users were mainly undergraduate students in the CSE department. There were 180 images used in the study. On average, each user searched for about 9 images. Each image was searched by six to seven different users.

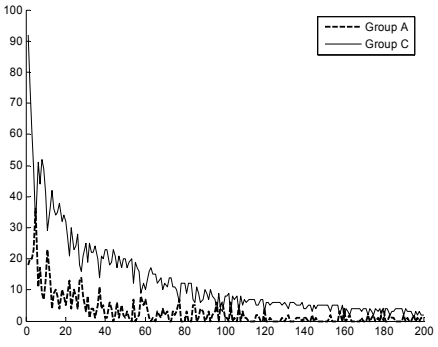
In this study, each user interacted with the term feedback interface to find target images through the web. To search for one target image, a user had to first provide an initial text query in his/her own terms. The system would then retrieve the top 20 images and show them to the user. Along with the retrieved images, the system also prompted 10 terms for the user to choose from. The user would be able to refine his or her query by selecting any terms from this list. This process was repeated until the user found a target or gave up the search. During the search, the system kept a log of the entire interaction. Based on the logged data, we collected 2483 terms (which appeared in the term list) with a total of 38655 occurrences. Each occurrence of these terms was assigned to the following four groups:

- Group A: the terms that are in the description of the target image and are also selected by the user. This group of terms are considered *informative* in the sense that if they occur in the description of the target image (i.e., if they are relevant to user's information needs), users can easily recognize these terms.
- Group B: the terms that are in the description of the target image but are not selected by the user. This group of terms are considered *uninformative* in the sense that even those terms appear in the target description (i.e., relevant), users have a hard time in recognizing them.
- Group C: the terms that are not in the description of the target image but are selected by the user. This group of terms are considered *overly-expressive* in the sense that even these terms are not relevant (i.e., do not occur in the target description), they tend to be mistakenly selected by a user.
- Group D: the terms that are not in the description of the target image and are not selected by the user. This group of terms are not particularly interesting for our purpose.

By separating terms into four groups, we were hoping to build a learning model to classify a term into three categories: informative, uninformative, and overly-expressive. Then based on such classification, each term could be given a weight indicating the priority of being included in the term list for users to choose. Ideally, "informative" terms should be given higher weights and "uninformative" and "overly-expressive" terms should be given lower weights.




(a) Frequencies of terms being "informative" and "uninformative"




(b) Frequencies of terms being "informative" and "overly-expressive"

**Figure 2: Term selection comparison**

Figure 2 shows the results from 200 most frequently occurred terms. Figure 2(a) shows the behavior of terms in Group A and Group B. We can see that almost every term appears more frequently in Group B than in Group A. In other words, almost every word appears more often as an uninformative term than as an informative one. This finding suggests that almost none of the terms can be categorized as "easily recognizable". Furthermore, the frequency ratio between "being uninformative" and "being informative" is steady across all terms. This indicates that no term seems to be significantly more informative than the other. Figure 2(b) shows the similar behavior between terms in Group A and terms in Group C. It indicates that almost every term is equally "confusing". For each term, on the one hand, when it occurs in the target description (i.e., relevant), it can be recognized. On the other hand, when it does not occur in the target description, it can also be mistakenly chosen. Two curves in Figure 2(b) also indicate that no term seems to be more "overly expressive" than the other.

As a conclusion, the empirical results from the second user study indicate that term classification based on terms themselves will not be successful. In fact terms by themselves do not show any strong indication as to how they will be perceived or recognized by a user. In the task of finding a target image, the content of the image and the image features

**Table 4. Measurement of term generation strategies**

| Strategy | *GenRate_des* at iteration 1 |
|---|---|
| 1: Random | 0.05 |
| 2: Freq_top | 0.15 |
| 3: Entropy_bottom | 0.17 |

play an important role on how a term is perceived. For example, the term 'tree' is intuitively an informative term. But if the tree objects are not in the prominent position in the target image, a strong response to the term "tree" will not be expected. Furthermore, different users may have different responses to a set of terms. A term may be informative to one user but not to the other. For example, the term "loch" may not mean anything to a user who is not familiar with this term. Therefore, our study indicates that term selection will not be improved only based on terms themselves. It requires more comprehensive study that takes into consideration of terms, visual features of images, and user cognitive models of perception.

## 4.2 Term Generation

The second direction is to identify good strategies to improve the term generation rate (i.e., increasing *GenRate_des* and decreasing *GenRate_undes*). Ideally, if the list of terms prompted to the user consists of more relevant terms (i.e., terms that appear in the description of a target image) and less irrelevant terms, then the chances for the desired terms to be included in the refined query will be improved.

We experimented with several strategies for term generation. Before describing different strategies, we first introduce some notations. Suppose our system retrieves 100 images based on a query. We use *Top Set* to denote the set of images that are shown to the user (i.e., the top retrieved images). If there are *N* images in the *Top Set*, then the next 100-*N* images will form the *Bottom Set* which are not shown to the user. In our experiments, the *Top Set* has the top 20 retrieved images.

Given these notations, we experimented with the following strategies to generate a list of terms prompted to the user:

- Strategy 1: Generate terms randomly from descriptions of those images in the *Top Set*. This provides a baseline.
- Strategy 2: Generate terms according to their frequencies of occurrence in the descriptions of images in the *Top Set*. This strategy is based on the hypothesis that the more a term appears in the *Top Set*, the more relevant the term relates to the searching topic, thus is more likely to describe the target image in the user's mind.
- Strategy 3: Generate terms based on the entropy of a term in the image descriptions in the *Bottom Set*. This strategy focuses on the ability of a term to separate the set of images that are not yet shown to the user, but could be related to the search topic. The stronger the ability is, the more efficient this term helps the system to narrow down the search space.

To evaluate these strategies, we conducted some experiments. In these experiments, totally 180 target images which have text descriptions were used. For each target image, we manually specified an initial query. Using this initial query, the system retrieved 100 images from the collection of 30,000 images. Based on the retrieved results, the system applied three different strategies to generate the term list. Since we did not want to mix in the effect of term selection, we only measured the desired term generation rate at the first iteration. The results are shown in Table 4. Since the third strategy shows slightly better performance, we used this one as the term generation strategy in our user study described in Section 3. Note that the results in Table 4 are the average *GenRate_des* at the first iteration, which is much higher than the average *GenRate_des* across all iterations (see Table 3). The reason is that, in our current design, we controlled the term generation so that terms generated will not be repeated in the following iterations. Therefore, the desired term generation rate diminishes as the iteration increases (as shown in Figure 3). Therefore, the average rate across all iterations (0.06 in Table 3) is much lower than the average rate at the first iteration in Figure 3.

We have also experimented with the inverse correlation strategy and the synonym strategy. The idea for the inverse correlation strategy is that if a term is strongly correlated with a query term given by a user, then that term carries less information in identifying new images that might be of user's interest. Therefore, we give a lower weight to the terms that are highly correlated with a query term using a vector space model. The idea for the synonym strategy is that if a term is the synonym of a query term, then it could be very relevant. We give it a higher weight since it maybe just a different vocabulary expressing the same meaning. To test this synonym strategy, we used WordNet. However, our current simulations have not shown the effectiveness of these two strategies in term generation.

The term generation study indicates that our current strategies only produce very low term generation rates. More strategies and approaches (e.g., techniques with pseudo relevance feedback) can be experimented in the future. Suppose term generation can be improved, the next question is what the

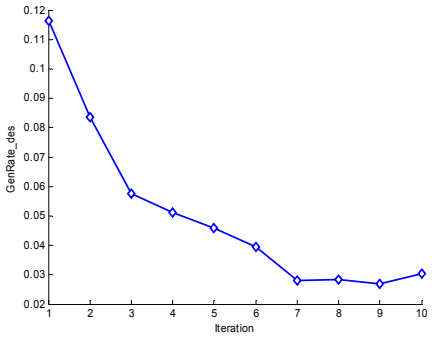



**Figure 3: The desired term generation rate diminishes as the iteration number increases**

impact of the improved term generation on the overall outlook of the term feedback interface? To what extend will the improved term generation rate affect the utility of the term feedback interface?

## 4.3 The Outlook

The motivation for the outlook investigation is to understand the potential of term feedback by taking into consideration technology limitations. If the technology for term generation were improved, what would be its impact on the utility of the term feedback interface? In other words, what is the requirement on term generation in order for term feedback to reach reasonable performance?

To study the potential outlook of the term feedback system, we examined the implications of different term generation rates on the overall retrieval performance. In particular, we conducted a simulation study using the fourteen images in our first user study (Section 3.2) (two out of the original sixteen images were removed because they did not have text descriptions). For each image, we had 8 initial queries collected from the first user study. Therefore, the simulation was repeated 8 times, each time with a different initial query. These formed a total of 112 (14*8) trials of image retrieval sessions.

For each trial, we simulated different term generation rates (i.e., *GenRate_des*) without actually applying any term generation algorithm. At each iteration after the initial query, a list of 10 terms is generated. In that list, (10**GenRate_des*) relevant terms were randomly picked from the description of the target image, and the rest irrelevant were generated according to their term frequencies in the *Top Set*. The system also simulated the user term selection rate. Our study on term selection (Section 4.1) indicates that the term selection rate will not be easily improved without more sophisticated studies. Therefore, our system only applied the empirical rate of *SelRate_des* and *SelRate_undes* (which are 0.26 and 0.08, respectively, from Table 3) to select the relevant/irrelevant terms from the term list to add to the query. The system then retrieved images based on the simulated queries. This retrieval process was repeated until a target image was retrieved or five iterations were reached.

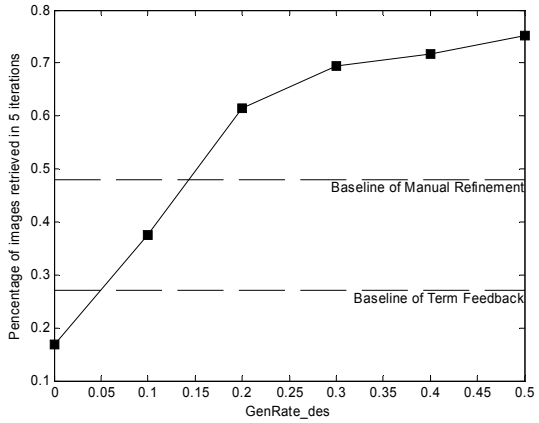



**Figure 4: The outlook of term feedback with respect to different desired term generation rates**

Figure 4 shows the simulation results. Here the X-axis represents different levels of the term generation rate (*GenRate_des*). The Y-axis represents the average percentage of target images that were successfully retrieved in five iterations among 112 trials. The curve in Figure 4 indicates that the final retrieval performance increases as the desired term generation rate increases. The two lines represent the performance of the term feedback interface and the manual refinement interfaces from the first user study (Table 1). As we can see that, the term feedback interface used in the first user study performs roughly at the same level as the generation rate 0.05 in the simulation. This is consistent with our previous results since the desired term generation rate for the term feedback interface used was 0.06 (see Table 3). The manual refinement interface performs roughly at the same level as the generation rate of 0.15. These results indicate that a considerable effort has to be made to improve the desired generation rate in order to make term feedback comparable to manual refinement. Thus, the effect of term feedback in interactive text-based image retrieval seems limited considering the performance of the current manual refinement techniques and the high requirement on term generation.

# 5. CONCLUSION

This paper presents our findings on the role of term feedback in interactive text-based image retrieval. The results from our user studies indicate that the term feedback interface seems not as effective as what we originally expected. Although it has a positive effect by allowing users to identify more relevant terms, it also has a strong negative effect by providing more opportunities for users to select irrelevant terms.

To understand factors that contribute to these effects and their implications on the potential of term feedback, we conducted further investigation. Our studies indicate that the retrieval performance using term feedback is dependent on two factors: (1) the term generation rate that applies to the underlying system mechanism in generating relevant terms as candidate terms in the list; and (2) the term selection rate that measures user responses to the prompted term list. It is shown that the term selection rate will be difficult to improve without more sophisticated studies that consider textual terms, image visual features, and user cognitive models of perceptions. It is also shown that, although the term generation rate could be improved, a considerable improvement has to be made in order to make the term feedback interface comparable to the current manual refinement interface. Therefore, term feedback may not be an effective strategy for our specific task in interactive text-based image retrieval.

We have also learned from this investigation that, in studying interactive technologies for information retrieval, it is desirable to apply both empirical user studies and simulation studies. While user studies can reveal user reactions and evaluate interactive systems, the simulation studies can provide a quick assessment on the potential utility before expensive implementation and user studies are taken place.

## 6. ACKNOWLEDGEMENT

This work was partially supported by grant IRGP-03-42111 from Michigan State University. The authors would like to thank Vineet Bansal for his contribution to the project and anonymous reviewers for their helpful comments and suggestions.

## 7. REFERENCES

[1] Anick, P. and Tipirneni, S. The Paraphrase Search Assistant: Terminological Feedback for Iterative Information Seeking. In *Proceedings of SIGIR'99*. Berkeley, CA, 1999.

[2] Belkin, N.J. and Marchetti, P.G. Determining the functionality and features of an intelligent interface to an information retrieval system, In *Proceedings of SIGIR'90*, Brussels, Belgium, 1990.

[3] Belkin, N.J., Cool, C., Koenemann, J., Ng, K.B., and Park, S. Using Relevance Feedback and Ranking in Interactive Searching. In *Proceedings of TREC4*. 1996.

[4] Blei, D. and Jordan, M. Modeling annotated data. In *Proceedings of 26th International Conference on Research and Development in Information Retrieval (SIGIR)*. 2003.

[5] Choi, Y. and Rasmussen, E. Search for Images: the Analysis of User's Queries for Image Retrieval in American History. In *Journal of the American Society for Information Science and Technology*, *54(6)*, April 2003.

[6] Clough, P., Sanderson, M., and Reid, Norman. The Eurovision St Andrews Photographic Collection. http://ir.shef.ac.uk/imageclef2004/guide.pdf.

[7] Harman, D. Towards Interactive Query Expansion. In *Proceeding of SIGIR'88*. 1988.

[8] He, X., King, O., Ma, W.-Y., Li, M., and Zhang, H. J. Learning a semantic space from user's relevance feedback for image retrieval. *IEEE Transaction on Curcuits and Systems for Video Technology*, 13(1):39-48, Jan. 2003

[9] Hearst, M. Using Categories to Provide Context for Full-Text Retrieval Results. In *Proceedings of RIAO'94.* 1994.

[10] Henniger, Scott and Nicholas, Belkin, Interface Issues and Interaction Strategies for Information Retrieval Systems. In *Proceedings of the Human Factors in computing Systems Conference (CHI'96)*, ACM Press, New York, 1996.

[11] Hoi, C.-H and Lyu, M. R. A Novel Log-based Relevance Feedback Technique in Content-based Image Retrieval, *Proc. of the 12th ACM International Conference on Multimedia.*

[12] Jeon, J., Lavrenko, V., and Manmatha, R. Automatic Image Annotation and Retrieval using Cross-Media Relevance Models. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*. 2003.

[13] Keister, L. H. User types and queries: impact on image access systems. In *Challenges in indexing electronic text and images (Fidel, R et al., eds)*. *ASIS*, 1994, 7-22

[14] Koenemann, J. and Belkin, N. J. A Case for Interaction: a Study of Interactive Information Retrieval Behavior and Effectiveness. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems: Common Ground*, p.205-212, April 13-18, 1996, Vancouver, Critish Columbia, Canada

[15] Lafferty, J. and Zhai, C. X. Document language models, query models, and risk minimization for information retrieval. In *The 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2001.

[16] Lavrenko, V. and Croft, B. Relevance-based langauge models. In *Proceeding of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2001.

[17] Lavrenko, V., Manmatha, R., and Jeon, J. A Model for Learning the Semantics of Pictures. In *Proceedings of Advance in Neutral Information Processing*. 2003.

[18] Li, J. and Wang, J.Z. Automatic linguistic indexing of pictures by a statistical modeling approach. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2003. 25(19): p. 1075-1088.

[19] McDonald, S. and Tait, J. Search Strategies in Content-based Imaged Retrieval. In *Proceedings of SIGIR 2003*. Toronto, Canada, 2003.

[20] Monay, F. and Gatica-Perez, D. On Image Auto-Annotation with Latent Space Models. In *Proc. ACM International Conference on Multimedia*. 2003.

[21] Ponte, J. *A Language Modeling Approach to Information Retrieval*, in Department of Computer Science, Univ. of Massachusetts at Amherst. 1998.

[22] Rui, Y., Huang, T. S., Ortega, M., and Mehrotra, S. Relevance Feedback: A Power Tool in Interactive Content-Based Image Retrieval, *IEEE Trans. on Circuits and Systems for Video Technology*, *Special Issue on Segmentation, Description, and Retrieval of Video Content*, pp644-655, Vol 8, No. 5, Sept, 1998.

[23] Santini, S. *Exploring image databases context-based retrieval.* Academic Press, San Diego USA, 2001.

[24] Vendrig, J., Worring, M., and Smeulders, A. Filter Image Browsing: Interactive Image Retrieval by Using Database Overviews. *Multimedial Tools and Applications*, 15, 83-103. 2001.