A Four-factor User Interaction Model for Content-Based Image Retrieval

Haiming Liu[1], Victoria Uren[1⋆], Dawei Song[2], Stefan Rüger[1]

[1]Knowledge Media Institute, The Open University, Milton Keynes, UK
[2]School of Computing, The Robert Gordon University, Aberdeen, UK
{h.liu,s.rueger}@open.ac.uk;v.uren@dcs.shef.ac.uk;d.song@rgu.ac.uk

Abstract. In order to bridge the "Semantic gap", a number of relevance feedback (RF) mechanisms have been applied to content-based image retrieval (CBIR). However current RF techniques in most existing CBIR systems still lack satisfactory user interaction although some work has been done to improve the interaction as well as the search accuracy. In this paper, we propose a four-factor user interaction model and investigate its effects on CBIR by an empirical evaluation. Whilst the model was developed for our research purposes, we believe the model could be adapted to any content-based search system.

Key words: User interaction, Relevance feedback, Content-based image retrieval

1 Introduction

Content-based image retrieval (CBIR) has been researched for decades, but it is not widely applied online. In our view, one of the reasons for this is that CBIR is normally performed by computing the dissimilarity between objects and queries based on their multidimensional feature vectors in content feature spaces, for example, colour, texture and structure features. There is a well known gap, called the "semantic gap", between the low-level feature of an image and its high-level meaning to users.

To help bridge this semantic gap, relevance feedback (RF) has been introduced into CBIR systems, which aims to bring users into the search loop. Existing research on RF [10] suggests that bringing users into the loop can help bridge the semantic gap and may also improve the retrieval accuracy. However, most existing RF techniques are highly system-centric. They focus more on improving search accuracy than the interaction between the system and users.

Therefore in an effort to develop more human-centric and user-oriented systems, Spink, et al. proposed a three-dimensional spatial model to support user interactive search for text retrieval [8]. The model emphasizes that partial relevance is as important as binary relevance/irrelevance, and indeed it can be more important for inexperienced users.

⋆ Present address: Department of Computer Science, Regent Court, 211 Portobello, University of Sheffield, Sheffield, S1 4DP United Kingdom.

Other existing research has been focused more on a single dimension, such as time. For example, Campbell in [1] proposed the Ostensive Model (OM) that indicates the degree of relevance relative to when a user selected the evidence from the results set. Later, Urban, et al. applied the so called increasing profile to CBIR [9]. Their preliminary study showed that the system based on the OM was preferred by users over traditional CBIR search engines.

Ruthven, et al. [6] adapted two dimensions from the Spink, et al model combined with OM in their study. Their experimental results showed that combining partial and time relevance did help the interaction between the user and the system.

Based on the related work, we are motivated to investigate what the outcome would be were we to combine the three-dimensional spatial model with the OM together and, further, to add another factor - frequency - to the combination. Therefore, in this paper, we propose an adaptive four-factor user interaction model (FFUIM) including relevance region, relevance level, time and frequency.

We will investigate the different interaction settings of the FFUIM, through simulated evaluations on a large image collection. The evaluation results provide initial evidence and insights into which interaction settings are likely to deliver the best search accuracy and lead to better user search experience.

2 User Interaction Models

In this section, we review a number of existing UI models and describe how our FFUIM harnesses their advantages, whilst addressing some of their limitations.

2.1 Three-dimensional spatial model

In order to improve the interaction between the users and the system, Spink, et al. proposed a three-dimensional spatial model of levels of relevance, regions of relevance and time of relevance to text retrieval [8]. Firstly, they applied Saracevic's five levels of relevance [7] as the way to indicate why the feedback is relevant, which confers a qualitative difference between levels. Secondly, the regions of relevance indicate the degree of users' relevance judgements to a feedback. The four regions are relevant, partially relevant, partially not relevant and not relevant. The third dimension is time of relevance, which is measured in formats such as information seeking stage and successive searches. We consider the model as a useful foundation from which to develop further user interaction models and techniques for CBIR.

2.2 Ostensive Model

Other research has tended to focus more on a single dimension, such as time. For example, Campbell in [1] proposed the Ostensive Model (OM) that indicates the degree of relevance relative to when a user selected the evidence from the results set. OM includes four ostensive relevance profiles: decreasing, increasing,

flat and current profiles, respectively. With the increasing profile the latest RF is deemed most important, whereas with the decreasing profile it is the earliest RF that is regarded as the most important. With the flat profile all RF is given equal importance, regardless of when the feedback was provided. Finally, the current profile gives the latest RF the highest weight and earlier RF is ignored. Campbell found that for text retrieval the increasing, flat and current profile showed overall better accuracy than the decreasing model, and the increasing profile was the most robust [1].

In [9] Urban et al. adapted the OM from text retrieval for CBIR to help overcome interaction problems between users and CBIR systems. In that study only the increasing profile was applied. The results indicated that, whilst users found the OM easy to use, they found it difficult to control the RF process without greater interaction. Furthermore, the traditional OM accepted only positive RF, whereas in reality users wish to refine their searches by providing both negative and positive RF. Indeed, some research [2,4,5] has shown that including negative examples into the RF can actually help improve the image retrieval accuracy.

2.3 Partial and ostensive evidence

Ruthven, et al. [6] adapted two dimensions from Spink, et al. model, namely: regions of relevance and time, for ranking query expansion terms in text retrieval. The region of relevance in their study is called partial evidence, which is a range of relevance level from one to ten. In addition, they applied the OM to the time dimension, which is called ostensive evidence. The ostensive evidence is measured by iterations of feedback. Their study shows that combining RF techniques with the user interaction factors is preferred by users over RF techniques alone. It will be interesting to see how the combined model performs in our CBIR system.

3 A Four-factor User Interaction Model for CBIR

Based on these interesting studies, we developed a new model named 'four-factor user interaction model (FFUIM)', which combines the three-dimensional spatial model with the OM and, further, to add another factor - frequency - to the combination. The FFUIM includes: relevance region, relevance level, time and frequency. We introduce the four factors in following sections.

3.1 Relevance Region

Instead of Spink, et al. four regions of relevance, the relevance region here comprises two parts: relevant (positive) evidence and non-relevant (negative) evidence. Both relevance regions contains a range of relevance levels.

3.2 Relevance Level

The relevance level here indicates how relevant/non-relevant the evidence is on the related relevance region, which implies a quantitative difference, and differs

from Saracevic's definition in Spink, et al. This factor is measured by a range of relevance level (integer 1-20) indicated by users. The distance function with the relevance level factor is given by

$$D_{ij} = d_{ij}/W_p, \tag{1}$$

where $D_{ij}(i = 1, 2, \ldots, m; j = 1, 2, \ldots, n)$ is the final distance between a query image i with an object image j; d_{ij} is the original distance between the query image i and an object image j; W_p is the partial weight, $W_p = r$ for the positive examples, and $W_p = \frac{1}{r}$ for the negative examples (r is the level of the relevance provided by the user between 1 and 20)[1] [2].

3.3 Time

We adapted the OM to the time factor to indicate the degree of relevance relative to when the evidence was selected. In this study, we have taken the OM a step further. In addition to using the increasing profile, we have also tested the flat profile, current profile and the decreasing profile. For our study, the increasing / decreasing profile means ostensive relevance weights for positive / negative examples increase / decrease respectively with further search iterations. The fundamental difference between our studies and Urban et al. is that we have applied these ostensive relevance weights to both the positive and negative feedback, and applied the weight to more than one image in every query. We propose the following distance function with ostensive weight:

$$D_{ij} = d_{ij}/W_o, \tag{2}$$

where W_o, the ostensive weight, can be different depending on the profile. $W_o = s$ for the positive examples, and $W_o = \frac{1}{s}$ for the negative examples (for the increasing profile, s is iterations of feedback; for the decreasing profile, s is iterations of feedback in the contrary order; for the flat profile, s is 1; for the current profile, s is 1 to current iteration, but 0 to previous iterations) [3].

3.4 Frequency

While we were investigating the combined models, we found that the same images can be used as positive/negative examples in different RF iterations. Thus, we wonder: can the number of times an image appears (frequency) across all the

[1] D_{ij} depending on positive d_{ij}/x and negative examples $d_{ij}/(1/x)$, but the later simplifies to $d_{ij} \times x$, here the x can be r,s,t. Therefore the distance become smaller the higher the positive weight and larger the higher the negative weight.

[2] Note that we have tested a number of other weighting functions for W_y (y can be o,p,f), e.g., $W_y = x$, $W_y = 2^x$ and $W_y = \ln(x)$ (x can be r,s,t) for positive examples, but there was no significant difference in performance (MAP). Here we use the linear setting for simplicity.

[3] Please see more detail in footnote 1 and 2.

iteration contribute to the model? To answer this question, we propose a new factor - frequency, which captures the number of appearances of an image in the user selected evidence both for positive and negative evidence separately. The distance function with frequency is given by

$$D_{ij} = d_{ij}/W_f, \tag{3}$$

where W_f, the frequency weight, is how often an image has been chosen as a relevant or non-relevant example: $W_f = t$ for the positive examples, and $W_f = \frac{1}{t}$ for the negative examples (t is the number of times the image was chosen as a feedback)[4].

4 Empirical Evaluation

Our empirical experiments aim to find possible interaction settings of the FFUIM that improve the search accuracy in comparison with a CBIR system without any interaction. The evaluation was a lab-based systematic comparison. We tested some individual and combined factors of the FFUIM. The performance indicator used was Mean Average Precision (MAP), and we used the ranking of images in the entire data set to compute the MAP for every experiment.

4.1 Experimental Setup

The ImageCLEFphoto2007 collection [3] was used, which consists of 20,000 real life images and 60 query topics. We applied colour feature HSV to all of the images. The City block distance (a special case of Minkowski distance family) was used to compute the distance between query images and object images.

Two Fusion Approaches. We used two fusion approaches to support two different RF scenarios. Firstly, the vector space model (VSM) [5] was deployed for positive relevance feedback only. By adding the weighting scheme of the FFUIM into the VSM, the approach is represented by:

$$D_{VSM} = \sum_i (d_{ij}/W_z), \tag{4}$$

where the D_{VSM} is the sum of the distance value between a query (containing i positive examples) and an object image j. W_z can be one of the three factors' weight W_o,W_p,W_f, or any combination weight of all three factors, depending upon which factor or combined factors is/are being tested.

Secondly, because the VSM in [5] only uses positive RF, we applied k-nearest neighbours (k-NN) for both positive and negative relevance feedback [5]. Here, by taking into account the weighting scheme, k-NN is given by:

[4] Please see more detail in footnote 1 and 2.

$$D_{KNN} = \frac{\sum_{i \in N}(d_{ij}/W_z + \varepsilon)^{-1}}{\sum_{i \in P}(d_{ij}/W_z + \varepsilon)^{-1} + \varepsilon}, \quad (5)$$

where D_{KNN} is the distance value between an object image j with all the example images (positive and negative) in the query. ε is a small positive number (e.g. 0.00001) to avoid division by zero. N and P denote the sets of positive and negative images in the query.

Two Interaction Approaches. Our experiments used two interaction approaches: pseudo RF and a method we call simulated user RF.

Firstly, pseudo RF was applied - a method widely used in information retrieval. Here there is no user interaction functionality with the RF approach. The system automatically takes the top three and bottom three images from the ranked last iteration search result of each query as positive and negative examples, respectively, to expand the current queries. The reason we take the bottom three images as negative feedback to expand the current queries is because, from our previous experiment, this approach outperforms the use of randomly chosen negative examples.

Secondly, so-called simulated user RF was used. This approach uses three truly relevant images from the top ranked results of each query and three non-relevant images from the bottom as tested against the official relevance judgments file. We derive this method to provide an automatic means of feedback which is closer to real user behavior. The reason we limit feedback to three positive images and three negative ones is because we want to make the experimental results more comparable with equal numbers of image examples in the queries.

For consistency of the two approaches, we used three image examples in each original query and each of the RF iterations. Further, we limited the number of iterations to be three, where iteration one is the search by original queries without RF, and iterations two and three are with RF. The time and relevance region factors are applied to all the queries on every iteration, whilst the relevance level and frequency factor is applied only to the latest query.

4.2 Experimental Results

Our experiment has tested the performance of 16 interaction settings of the FFUIM, which includes four profiles of OM (time factor): flat profile, increasing profile, current profile, decreasing profile, these profiles combined with the relevance level factor, and the above combinations joint with the frequency factor. Each of the 16 settings was tested using positive RF only as well as positive and negative RF (relevance region factor). The models have been tested against a large image collections and two interaction approaches as previously described. The following insights and analysis has been made, by doing statistical significance tests (the Wilcoxon signed ranks test with $\alpha = 0.05$):

Firstly, simulated user RF has better performance than pseudo RF. Secondly, with the pseudo RF approach, accuracy falls with increasing iterations. Thirdly,

under simulated user RF approach, the performance clearly improves with each search iteration for all the results.

Apart from these generic insights, other results vary depending on the different settings and iterations. Since iteration three is the last iteration in our experiment and the weights should show more effect on the results, and, in addition, the simulated user RF outperforms pseudo RF and is closer to the real search scenario, we have undertaken further detailed analysis of the simulated RF at iteration three based on different search settings as follows:

Comparing the four profiles of the Ostensive Model (time factor). For the positive examples only setting, the decreasing and current profiles show consistently good performance, then the flat profile outperforms the increasing profile in most tests; for the both positive and negative example setting, the decreasing, flat and increasing profiles are not significantly different, but the current profile shows statistically worse performance than the other three profiles. The results do not show the same observation as previous OM studies, namely that the latest RF expresses best the user's information needs. This may be because the relevance judgement file was developed against the original query that is the oldest RF iteration. Thus the decreasing profile performs consistently well in different circumstances. These models need further testing in a real as opposed to simulated CBIR search environment.

With or without relevance level factor. In all of the tests, the relevance level when combined with the OM is not significantly different to the OM alone. This factor also needs further testing under a real user as opposed to simulated user evaluation.

With or without frequency factor. The frequency factor when combined with the other factors does not lead to significantly better performance than the factors without frequency factor. This may be because the limited number of search iterations means that the frequency weight has little impact. In addition, our definition of the frequency factor is that the latest query images are more important, which is different from the relevance judgement file that was created based on the original queries. This result may be clearer when we run further iterations of the experiment, or even under a real as opposed to simulated user evaluation.

Positive examples only and both positive and negative examples (relevance region factor). The use of both positive and negative example RF with k-NN approach performs significantly better than only positive example RF with VSM approach. The promising result encourages us to include the negative functionalities to our future visual search system, and then we need to think about how to deliver these functionalities to users through the interface.

5 Conclusion and Future Work

In an effort to alleviate the limitations of current user interaction models and to find a UI model to deliver a better interaction and search accuracy for CBIR, we have proposed a new four-factor user interaction model based on relevance

region, relevance level, time and frequency. We have also empirically investigated different settings of the proposed model.

The following main observations have been made from the evaluation results: (1) bringing the user into the loop will enhance CBIR; (2) allowing both positive and negative feedback improves search performance; (3) combining the relevance level and frequency factor with other factors will make the user interaction model more usable and may well improve the search accuracy.

This work will be a foundation for developing more effective user interaction systems for CBIR. We have developed a visual content-based image search system, so that we can carry out real as opposite to simulated user experiments to evaluate the usefulness and effectiveness of the different settings of the FFUIM model. We are using a series of quantitative performance indicators, such as scores from questionnaires, precision of actual search results, time and number of clicks taken to complete the task, etc. Early results of the user study are under review and detailed analysis is underway.

References

1. I. Campbell. Interactive evaluation of the ostensive model using a new test collection of images with multiple relevance assessments. *Journal of Information Retrieval*, 2(1), 2000.
2. M. D. Dunlop. The effect of accessing nonmatching documents on relevance feedback. *ACM Transactions on Information Systems (TOIS)*, 15(2):137–153, 1997.
3. M. Grubinger, P. Clough, H. Müller, and T. Deselaers. The iapr tc-12 benchmark: A new evaluation resource for visual information systems. In *In Proceedings of International Workshop OntoImage2006 Language Resources for Content-Based Image Retrieval*, pages 13–23, 2006.
4. H. Müller, W. Müller, S. Marchand-Maillet, and T. Pun. Strategies for positive and negative relevance feedback in image retrieval. In *Proceedings of the International Conference on Pattern Recognition (ICPR'2000)*, volume 1, pages 1043–1046, Barcelona, Spain, September 2000.
5. M. J. Pickering and S. Rüger. Evaluation of key frame-based retrieval techniques for video. *Computer Vision and Image Understanding*, 92(2-3):217–235, November 2003.
6. I. Ruthven, M. Lalmas, and K. van Rijsbergen. Incorporating user search behaviour into relevance feedback. *Journal of the American Society for Information Science and Technology*, 54(6):528–548, 2003.
7. T. Saracevic. Relevance reconsidered. In *Proceedings of the Second Conference on Conceptions of Library and Information Science (CoLIS 2)*, pages 210–218, Copenhagen,Denmark, October 1996.
8. A. Spink, H. Greisdorf, and J. Bateman. From highly relevant to not relevant: examining different regions of relevance. *Information Processing Management*, 34(5):599–621, 1998.
9. J. Urban, J. M. Jose, and K. van Rijsbergen. An adaptive technique for content-based image retrieval. *Multimedia Tools and Applications*, 31:1–28, July 2006.
10. X. S. Zhou and T. S. Huang. Relevance feedback in image retrieval: A comprehensive review. *Multimedia Systems*, 8(6):536–544, April 2003.