Multi–modal relevance feedback for medical image retrieval

Dimitrios Markonis
HES–SO
TechnoPole 3
Sierre, Switzerland
dimitrios.markonis@hevs.ch

Roger Schaer
HES–SO
TechnoPole 3
Sierre, Switzerland
roger.schaer@hevs.ch

Henning Müller
HES–SO
TechnoPole 3
Sierre, Switzerland
henning.mueller@hevs.ch

ABSTRACT

Medical image retrieval can assist physicians in finding information supporting their diagnosis. Systems that allow searching for medical images need to provide tools for quick and easy navigation and query refinement as the time for information search is often short.

Relevance feedback is a powerful tool in information retrieval. This study evaluates relevance feedback techniques with regard to the content they use. A novel relevance feedback technique that uses both text and visual information of the results is proposed.

Results show the potential of relevance feedback techniques in medical image retrieval and the superiority of the proposed algorithm over commonly used approaches.

Future steps include integrating semantics into relevance feedback techniques to benefit of the structured knowledge of ontologies and experimenting on the fusion of text and visual information.

Keywords

relevance feedback, content–based image retrieval, medical image retrieval

1. INTRODUCTION

Searching for images is a daily task for many medical professionals, especially in image–oriented fields such as radiology. However, the huge amount of visual data in hospitals and the medical literature is not always easily accessible and physicians have generally little time for information search as they are charged with many tasks.

Therefore, medical image retrieval systems need to return information adjusted to the knowledge level and expertise of the user in a quick and precise fashion. A well known technique trying to improve search results by user interaction is relevance feedback [13]. Relevance feedback allows the user to mark results returned in a previous search step as relevant or irrelevant to refine the initial query. The concept behind relevance feedback is that though user may have difficulties in formulating a precise query for a specific task, they generally see quickly whether a returned result is relevant to the information need or not. This technique found use in image retrieval particularly with the emerge of content–based image retrieval (CBIR) systems [18, 19, 20]. Following the CBIR mentality, the visual content of the marked results is used to refine the initial image query. With the result images represented as a grid of thumbnails, relevance feedback can be applied quickly to speed up the search iterations and refine results. Recent user–tests with radiologists on a medical image search system also showed that this method is intuitive and straightforward to learn [7].

Depending on whether the user manually provides the feedback to the system (e.g. by marking results) or the system obtains this information automatically (e.g. by log analysis) relevance feedback can be categorized as explicit or implicit. Moreover, the information obtained by relevance feedback can be used to affect the general behaviour of the system (long–term learning). In [11] a market basket analysis algorithm is applied in image retrieval of long–term learning. A recent review of short–term and long–term learning relevance feedback techniques in CBIR can be found in [6]. An extensive survey of relevance feedback in text–based retrieval systems is presented in [15] and for CBIR in [14].

In the medical informatics field, [1] applies CBIR with relevance feedback on mammography retrieval. In [12], an image retrieval framework using relevance feedback is evaluated on a dataset of 5000 medical images that uses support vector machines to compute the refined queries.

In this paper we evaluate different explicit, short–term relevance feedback techniques using visual content or text for medical image retrieval. We propose a technique that combines visual and text–based relevance feedback and show that it achieves a competitive performance to the state–of–the–art approaches.

2. METHODS

2.1 Rocchio algorithm

One of the most well known relevance feedback techniques is Rocchio's algorithm [13]. Its mathematical definition is given below:

$$\vec{q}_m = \alpha\vec{q}_o + \beta\frac{1}{|D_r|}\sum_{\vec{d}_j \in D_r}\vec{d}_j - \gamma\frac{1}{|D_{nr}|}\sum_{\vec{d}_j \in D_{nr}}\vec{d}_j \qquad (1)$$

where $\vec{q}_m$ is the modified query,
$\vec{q}_o$ is the original query,
D_r is the set of relevant images,

MedIR 2014, July 11, 2014, Gold Coast, Australia.

D_{nr} is the set of non–relevant images and
α, β and γ are weights.

Typical values for the weights are $\alpha = 1, \beta = 0.8$ and $\gamma = 0.2$. Rocchio's algorithm is typically used in vector models and also for CBIR. Intuitively, the original query vector is moved towards the relevant vectors and away from the irrelevant ones. By giving a weight to the positive and negative parts a problem of CBIR can be avoided that when more negative than positive feedback exists that also many relevant images disappear from the results set.

2.2 Late fusion

Another technique that showed potential in image retrieval [5] is late fusion. Late fusion [2] is used in information retrieval to combine result lists. It can be applied for fusing multiple features, multiple queries and in multi–modal techniques. The concept behind this method is to merge the result lists into a single list while boosting common occurrences using a fusion rule.

For example, the fusion rule of the score–based late fusion method CombMNZ [17] is defined as:

$$S_{\texttt{combMNZ}}(i) = F(i) * S_{\texttt{combSUM}}(i) \tag{2}$$

where $F(i)$ is the number of times an image i is present in retrieved lists with a non–zero score, and $S(i)$ is the score assigned to image i. CombSUM is given by

$$S_{\texttt{combSUM}}(i) = \sum_{j=1}^{N_j} S_j(i) \tag{3}$$

where $S_j(i)$ is the score assigned to image i in retrieved list j.

2.3 Multi–modal relevance feedback

Most of the techniques use vectors either from the text or the visual models. However, it has been shown that approaches that use both text and visual information can outperform single–modal ones in image retrieval. We propose the use of multi–modal information for relevance feedback to enhance the retrieval performance. This is, to the extend of our knowledge, the first time that such a technique is proposed in image retrieval. As late fusion is applied on result lists, it is straightforward to use for combining results from visual and text queries.

2.4 Experimental setup

For evaluating the relevance feedback techniques the following experimental setup was followed: The n search iterations are initiated with a text query in iteration 0. The relevant results from the top k results of iteration i were used in the relevance feedback formula of the iteration $i + 1$ for $i = 0 ... n - 2$.

The image dataset, topics and ground truth of ImageCLEF 2012 medical image retrieval task [9] were used in this evaluation. The dataset contains more than 300'000 images from the medical open access literature.

The image captions were accessed by the text–based runs and indexed with the Lucene[1] text search engine. Vector space model was used along with tokenization, stopword removal, stemming and Inverse document frequency-Term frequency weighting. The Bag–of–visual–words model described in [3] and the bag–of–colors model appearing in [4] were used for the visual modelling of the images. In multimodal runs, the fusion of the visual and text information is performed only for the text 1000 top results as in the evaluation of ImageCLEF only the top 1000 documents are taken into account in any case.

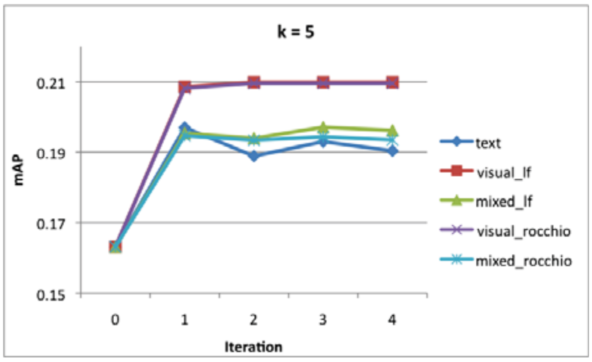

Figure 1: Mean average precision per search iteration for $k = 5$.

Table 1: Best mAP scores

Run	k = 5	k = 20	k = 50	k = 100
text	0.197 (1)	0.2544 (4)	0.3107 (3)	0.3349 (4)
visual_lf	0.2099 (2)	0.2243 (3)	0.2405 (4)	0.2553 (3)
visual_roc	0.2096 (2)	0.2187 (2)	0.2249 (3)	0.2268 (2)
mixed_lf	0.1971 (3)	0.2606 (4)	0.3079 (4)	0.3487 (3)
mixed_roc	0.1947 (1)	0.2635 (4)	0.3207 (4)	0.3466 (4)

Five techniques were evaluated in this study:

1. **text**: text–based RF using vector space model. Word stemming, tokenization and stopword removal is performed in both text and multi–modal runs.

2. **visual_rocchio**: visual RF using Rocchio to fuse the relevant image vectors and CombMNZ fusion to fuse the original query's results with the visual ones.

3. **visual_lf**: visual RF using late fusion (and the CombMNZ fusion rule) to fuse the relevant image results and the original query results with the visual ones.

4. **mixed_rocchio**: multimodal RF using Rocchio to fuse the relevant image vectors and CombMNZ fusion to fuse the original query results with the relevant caption results and relevant visual results.

5. **mixed_lf**: multimodal RF using late fusion (and the CombMNZ fusion rule) to fuse the relevant image results and the original query results with the captions' results and relevant visual results.

3. RESULTS

The evaluation of the five techniques was performed for $k = 5, 20, 50, 100$ and $n = 5$. Results of the mean average precision (mAP) of each technique per iteration are shown in Figures 1, 2, 3, 4.

Table 1 gives the best mAP scores of each run. The numbers in parentheses are the number of the iteration when this score was achieved. For scores that were the same in multiple iterations of the same run, the iteration closer to the first is used.

[1] http://lucene.apache.org/

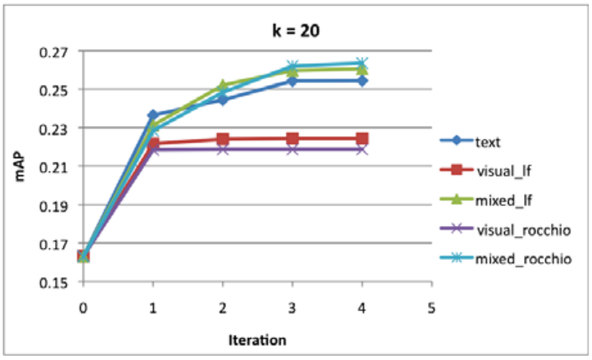

Figure 2: Mean average precision per search iteration for $k = 20$.

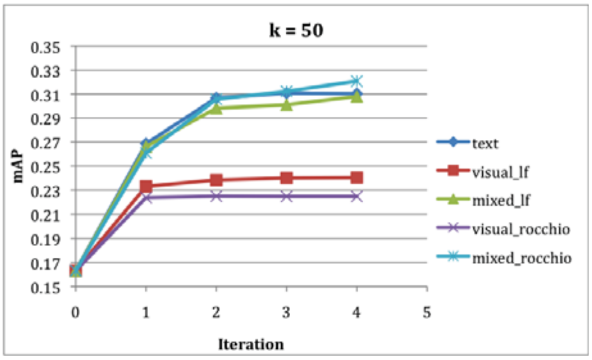

Figure 3: Mean average precision per search iteration for $k = 50$.

4. DISCUSSION

All of the evaluated techniques improve retrieval after the initial search iteration. This demonstrates the potential of relevance feedback for refining medical image search queries.

Relevance feedback using only visual appearance models, even though improving the retrieval performance after the first iteration, performed worse than the text–based runs in most cases. Visual features still suffer from the semantic gap between the expressiveness of visual features and our human interpretation. Still, this shows their usefulness in image datasets where no or little text meta–data are available. Moreover, when combined with the text–information in the proposed method, they improve the text–only baseline.

The proposed multi–modal runs provide the best results in all the cases except for case $k = 5$. Surprisingly, the visual runs perform slightly better than the text and the multi–modal approaches for this case. However, assuming independent and normal distributed average precision values the significance tests show that the difference is not statistically significant.

We consider the case $k = 20$ as the most realistic scenario since users do not often inspect more than 2 pages of results. Especially for grid–like result interface views, where each page can contain 20 to 50 results, we consider $k = 20$ more realistic than $k = 5$. In this case the proposed methods achieve the best performance with 0.2606 and 0.2635 respectively. Again, the significance tests do not find any significance difference between the three best approaches. However, applying different fusion rules for combining visual and text information (such as linear–weighting) could further improve the results of the mixed approaches.

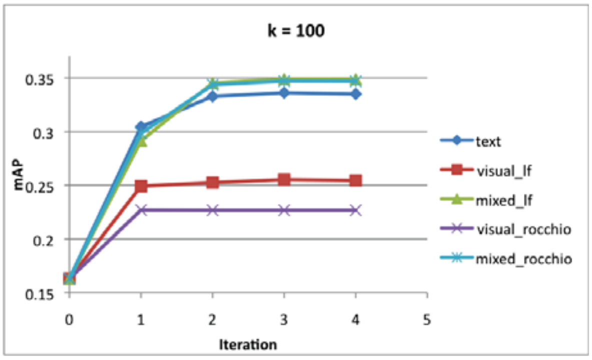

Figure 4: Mean average precision per search iteration for $k = 100$.

It can be noted that as the k increases, the performance improvement also increases, highlighting the added value of relevance feedback. Larger values of k were not explored as this scenarios were judged as unrealistic.

In the visual runs using Rocchio for combining the visual queries is performing worse than late fusion. This comes in accordance with the findings in [3]. The reason behind this could be that the large visual diversity of relevant images in medicine and the curse of dimensionality cause the modified vector to behave as an outlier in the high dimensional visual feature space. In the mixed runs the difference between the two methods is not statistically significant with Rocchio performing slightly better than the late fusion.

Irrelevant results were ignored, as they often have little or no impact on the retrieval performance [10, 16]. More importantly, the ground truth of the dataset used contains a much larger portion of annotated irrelevant results than relevant ones. This was considered to potentially simulate an unrealistic scenario, as users do not usually mark many results as negative examples. Having too many negative examples could also cause the modified vector to follow an outlier behaviour. Preliminary results confirmed this hypothesis, where the use of negative results for relevance feedback can decrease performance after the first iteration.

It should be noted that this is an automated relevance feedback experiment of positive only feedback and that in selective relevance feedback situations the retrieval performance is expected to perform even better. A larger number of steps could be investigated but this might be unrealistic, given the fact that physicians have little time and stop after a few minutes of search [8]. Often users will only test a few steps of relevance feedback at the most.

5. CONCLUSIONS

This paper proposes the use of multi–modal information when applying relevance feedback to medical image retrieval. An experiment was set up to simulate the relevance feedback of a user on a number of medicine–related topics from ImageCLEF 2012.

In general, all the techniques evaluated in this study improve the performance, which shows the added value of relevance feedback. Text–based relevance feedback showed consistently good results. Visual–based techniques showed competitive performance for small shortlist sizes, underperforming in the rest of the cases. The proposed multi–modal approaches showed promising results slightly outperforming the text–based one but without statistical significance.

More fusion techniques are going to be evaluated in the future. Comparison to manual query refinement by users is considered in future plans, to assess relevance feedback as a concept in medical image retrieval. The addition of semantic search is also of interest, to take advantage of the structured knowledge of the medical ontologies such as RadLex (Radiology Lexicon) and MeSH (Medical Subject Headings).

6. ACKNOWLEDGEMENTS

This work was supported by the EU 7th Framework Program in the context of the Khresmoi project (grant 257528).

7. REFERENCES

[1] C.-C. Chen, P.-J. Huang, C.-Y. Gwo, Y. Li, and C.-H. Wei. Mammogram retrieval: Image selection strategy of relevance feedback for locating similar lesions. *International Journal of Digital Library Systems (IJDLS)*, 2(4):45–53, 2011.

[2] A. Depeursinge and H. Müller. Fusion techniques for combining textual and visual information retrieval. In H. Müller, P. Clough, T. Deselaers, and B. Caputo, editors, *ImageCLEF*, volume 32 of *The Springer International Series On Information Retrieval*, pages 95–114. Springer Berlin Heidelberg, 2010.

[3] A. García Seco de Herrera, D. Markonis, I. Eggel, and H. Müller. The medGIFT group in ImageCLEFmed 2012. In *Working Notes of CLEF 2012*, 2012.

[4] A. García Seco de Herrera, D. Markonis, and H. Müller. Bag of colors for biomedical document image classification. In H. Greenspan and H. Müller, editors, *Medical Content–based Retrieval for Clinical Decision Support*, MCBR–CDS 2012, pages 110–121. Lecture Notes in Computer Sciences (LNCS), Oct. 2013.

[5] A. García Seco de Herrera, D. Markonis, R. Schaer, I. Eggel, and H. Müller. The medGIFT group in ImageCLEFmed 2013. In *Working Notes of CLEF 2013 (Cross Language Evaluation Forum)*, September 2013.

[6] J. Li and N. M. Allinson. Relevance feedback in content-based image retrieval: a survey. In *Handbook on Neural Information Processing*, pages 433–469. Springer, 2013.

[7] D. Markonis, F. Baroz, R. L. Ruiz de Castaneda, C. Boyer, and H. Müller. User tests for assessing a medical image retrieval system: A pilot study. In *MEDINFO 2013*, 2013.

[8] D. Markonis, M. Holzer, S. Dungs, A. Vargas, G. Langs, S. Kriewel, and H. Müller. A survey on visual information search behavior and requirements of radiologists. *Methods of Information in Medicine*, 51(6):539–548, 2012.

[9] H. Müller, A. García Seco de Herrera, J. Kalpathy-Cramer, D. Demner Fushman, S. Antani, and I. Eggel. Overview of the ImageCLEF 2012 medical image retrieval and classification tasks. In *Working Notes of CLEF 2012 (Cross Language Evaluation Forum)*, September 2012.

[10] H. Müller, W. Müller, D. M. Squire, S. Marchand-Maillet, and T. Pun. Strategies for positive and negative relevance feedback in image retrieval. Technical Report 00.01, Computer Vision Group, Computing Centre, University of Geneva, rue G n ral Dufour, 24, CH–1211 Gen ve, Switzerland, Jan. 2000.

[11] H. Müller, D. M. Squire, and T. Pun. Learning from user behavior in image retrieval: Application of the market basket analysis. *International Journal of Computer Vision*, 56(1–2):65–77, 2004. (Special Issue on Content–Based Image Retrieval).

[12] M. M. Rahman, P. Bhattacharya, and B. C. Desai. A framework for medical image retrieval using machine learning and statistical similarity matching techniques with relevance feedback. *Information Technology in Biomedicine, IEEE Transactions on*, 11(1):58–69, 2007.

[13] J. J. Rocchio. Relevance feedback in information retrieval. In *The SMART Retrieval System, Experiments in Automatic Document Processing*, pages 313–323. Prentice Hall, Englewood Cliffs, New Jersey, USA, 1971.

[14] Y. Rui, T. S. Huang, and S. Mehrotra. Relevance feedback techniques in interactive content–based image retrieval. In I. K. Sethi and R. C. Jain, editors, *Storage and Retrieval for Image and Video Databases VI*, volume 3312 of *SPIEProc*, pages 25–36, Dec. 1997.

[15] I. Ruthven and M. Lalmas. A survey on the use of relevance feedback for information access systems. *The Knowledge Engineering Review*, 18(02):95–145, 2003.

[16] G. Salton and C. Buckley. Improving retrieval performance by relevance feedback. *Readings in information retrieval*, 24:5, 1997.

[17] J. A. Shaw and E. A. Fox. Combination of multiple searches. In *TREC-2: The Second Text REtrieval Conference*, pages 243–252, 1994.

[18] D. M. Squire, W. Müller, H. Müller, and T. Pun. Content–based query of image databases: inspirations from text retrieval. *Pattern Recognition Letters (Selected Papers from The 11th Scandinavian Conference on Image Analysis SCIA '99)*, 21(13–14):1193–1198, 2000. B.K. Ersboll, P. Johansen, Eds.

[19] L. Taycher, M. L. Cascia, and S. Sclaroff. Image digestion and relevance feedback in the ImageRover WWW search engine. pages 85–94, 1997.

[20] M. E. Wood, N. W. Campbell, and B. T. Thomas. Iterative refinement by relevance feedback in content–based digital image retrieval. pages 13–20, 1998.