Manual Query Modification and Data Fusion for Medical Image Retrieval

Jeffery R. Jensen and William R. Hersh

Department of Medical Informatics & Clinical Epidemiology
Oregon Health & Science University
Portland, Oregon, USA
{jensejef, hersh}@ohsu.edu

Abstract. Image retrieval has great potential for a variety of tasks in medicine but is currently underdeveloped. For the ImageCLEF 2005 medical task, we used a text retrieval system as the foundation of our experiments to assess retrieval of images from the test collection. We conducted experiments using automatic queries, manual queries, and manual queries augmented with results from visual queries. The best performance was obtained from manual modification of queries. The combination of manual and visual retrieval results resulted in lower performance based on mean average precision but higher precision within the top 30 results. Further research is needed not only to sort out the relative benefit of textual and visual methods in image retrieval but also to determine which performance measures are most relevant to the operational setting.

1 Introduction

The goal of the medical image retrieval task of ImageCLEF is to identify and develop methods to enhance the retrieval of images based on real-world topics that a user would bring to such an image retrieval system. A test collection of nearly 50,000 images - annotated in English, French, and/or German - and 25 topics provided the basis for experiments. As described in the track overview paper [1], the test collection was organized from four collections, each of which was organized into cases consisting of one or more images plus annotations at the case or image level (depending on the organization of the original collection).

There are two general approaches to image retrieval, semantic (also called context-based) and visual (also called content-based) [2]. Semantic image retrieval uses textual information to determine an image's subject matter, such as an annotation or more structured metadata. Visual image retrieval, on the other hand, uses features from the image, such as color, texture, shapes, etc., to determine its content. The latter has historically been a difficult task, especially in the medical domain [3]. The most success for visual retrieval has come from "more images like this one" types of queries. There has actually been little research in the types of techniques that would achieve good performance for queries more akin to those a user might enter into a text retrieval system, such as "images showing different types of skin cancers." Some

C. Peters et al. (Eds.): CLEF 2005, LNCS 4022, pp. 673 – 679, 2006.

researchers have begun to investigate hybrid methods that combine both image context and content for indexing and retrieval [3].

Oregon Health & Science University (OHSU) participated in the medical image retrieval task of ImageCLEF 2005. Our experiments were based on a semantic image retrieval system, although we also attempted to improve our performance by fusing our results with output from a content-based search. Data fusion has been used for a variety of tasks in IR, e.g., [4]. Our experimental runs included an automatic query, a manually modified query, and a manual/visual query (the manual query refined with the results of a visual search).

2 System Overview

Our retrieval system was based on the open-source search engine, Lucene. We have used Lucene in other retrieval evaluation forums, such as the Text Retrieval Conference (TREC) Genomics Track [5]. Documents in Lucene are indexed by parsing of individual words and weighting of those words with an algorithm that sums for each query term in each document the product of the term frequency (TF), the inverse document frequency (IDF), the boost factor of the term, the normalization of the document, the fraction of query terms in the document, and the normalization of the weight of the query terms, for each term in the query. The score of document d for query q consisting of terms t is calculated as follows:

$$score(q,d) = \sum_{t\ in\ q} tf(t,d) * idf(t) * boost(t,d) * norm(d,t) * frac(t,d) * norm(q)$$

where: tf(t.d) = term frequency of term t in document d

idf(t) = inverse document frequency of term t
boost(t,d) = boost for term t in document d
norm(t,d) = normalization of d with respect to t
frac(t,d) = fraction of t contained in d
norm(q) = normalization of query q

Lucene is distributed with a variety of analyzers for textual indexing. We chose Lucene's standard analyzer, which supports acronyms, floating point numbers, lowercasing, and stop word removal. The standard analyzer was chosen to bolster precision. Each annotation, within the library, was indexed with three data fields, which consisted of a collection name, a file name, and the contents of the file to be indexed. Although the annotations were structured in XML, we indexed each annotation without the use of an XML parser. Therefore, every XML element was indexed (including its tag) along with its corresponding value.

As noted in the track overview paper, some images were indexed at the case level, i.e., the annotation applied to all images associated with the case. (This applied for the Casimage and MIR collections, but not the PEIR or PathoPIC collections.) When the search engine matched a case annotation, each of the images associated with the case was added to the retrieval output. It was for this reason that we also did a run that filtered the output based on retrieval by a visual retrieval run, in an attempt to focus the output of images by whole cases.

3 Methods

OHSU submitted three official runs for ImageCLEF 2005 medical image retrieval track. These included two that were purely semantic, and one that employed a combination of semantic and visual searching methods.

Our first run (OHSUauto) was purely semantic. This run was a "baseline" run, just using the text in the topics as provided with the unmodified Lucene system. Therefore, we used the French and German translations that were also provided with the topics. For our ranked image output, we used all of the images associated with each retrieved annotation.

For our second run (OHSUman), we carried out manual modification of the query for each topic. For some topics, the keywords were expanded or mapped to more specific terms. This made the search statements for this run more specific. For example, one topic focused on chest x-rays showing an enlarged heart, so we added a term like cardiomegaly. Since the manual modification resulted in no longer having accurate translations, we "expanded" the manual queries with translations that were obtained from Babelfish (http://babelfish.altavista.com). The newly translated terms were added to the query with the text of each language group (English, German, and French) connecting via a union (logical OR). Figure 1 shows a sample query from this run.

In addition to the minimal term mapping and/or expansion, we also increased the significance for a group of relevant terms using Lucene's "term boosting" function. For example, for the topic focusing on chest x-rays showing an enlarged heart; we increased the significance of documents that contained the terms, chest and x-ray and posteroanterior and cardiomegaly, while the default significance was used for documents that contained the terms, chest or x-ray or posteroanterior, or cardiomegaly. This strategy was designed to give a higher rank to the more relevant documents within a given search. Moreover, this approach attempted to improve the precision of the results from our first run. Similar to the OHSUauto run, we returned all images associated with the retrieved annotation.

```
(AP^2 PA^2 anteroposterior^2 posteroanterior^2 thoracic thorax cardiomegaly^3
heart coeur)
```

Fig. 1. Manual query for topic 12

Our third run (OHSUmanviz) used a combination of textual and visual retrieval methods. We took the image output from OHSUman and excluded all documents that were not retrieved by the University of Geneva "baseline" visual run (GE_M_4g.txt). In other words, we performed an intersection (logical AND) between the OHSUman and GE_M_4g.txt runs as a "combined" visual and semantic run.

Consistent with the ImageCLEF medical protocol, we used mean average precision (MAP) as our primary outcome measure. However, we also analyzed other measures output from trec_eval, in particular the precision at N images measures.

4 Results

Our automatic query run (OHSUauto) had the largest number of images returned for each topic. The MAP for this run was extremely low at 0.0403, which fell below the median (0.11) of the 9 submissions in the “text-only automated” category.

The manually modified queries run (OHSUman) for the most part returned large numbers of images. However, there were some topics for which it returned fewer images than the OHSUauto run. Two of these topics were those that focused on Alzheimer’s disease and hand-drawn images of a person. This run was in the “text-only manual” category and achieved an MAP of 0.2116. Despite being the only submission in this category, this run scored above any run from the “text-only automatic” category and as such was the best text-only run.

When we incorporated visual retrieval data (OHSUmanviz), our queries returned the smallest number images for each topic. The intent was to improve precision of the results from the previous two techniques. This run was placed in the “text and visual manual” category, and achieved an MAP of 0.1601, which was the highest score in this category. This technique’s performance was less than that of our manual query technique. Recall that both our manual and manual/visual techniques used the same textual queries, so the difference in the overall score was a result of the visual refinement.

Figure 2 illustrates the number of images returned by each of the techniques, while Figure 3 shows MAP per topic for each run. Even though the fully automatic query technique consistently returned the largest number of images on a per-query basis, this approach rarely outperformed the others. Whereas the manual query technique did not consistently return large numbers of images for each query, it did return a good proportion of relevant images for each query. The manual/visual query technique found a good proportion of relevant images but clearly eliminated some images that the text-only search found, resulting in decreased performance.

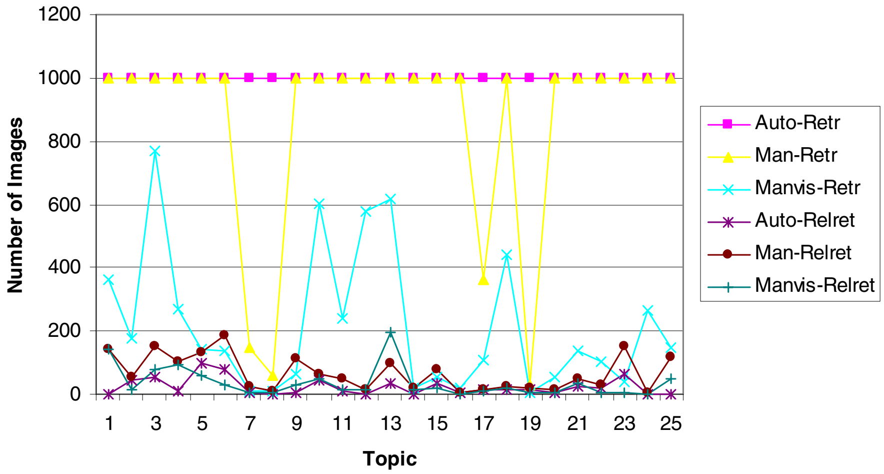

Fig. 2. Number of relevant images and retrieved relevant images for each of the three runs for each topic

Perhaps the most interesting result from all of our runs was comparing the performance based on MAP with precision at top of the output. Despite the overall lower MAP, the OHSUmanvis had better precision starting at five images and continuing to 30 images. The better MAP is explainable by the high precision across the remainder of the output (down to 1,000 images). However, this finding is significant by virtue of the fact that many real-world uses of image retrieval may have users who explore output solely in this range. Figure 4 shows precision at various levels of output, while Figure 5 shows a recall-precision curve comparing the two.

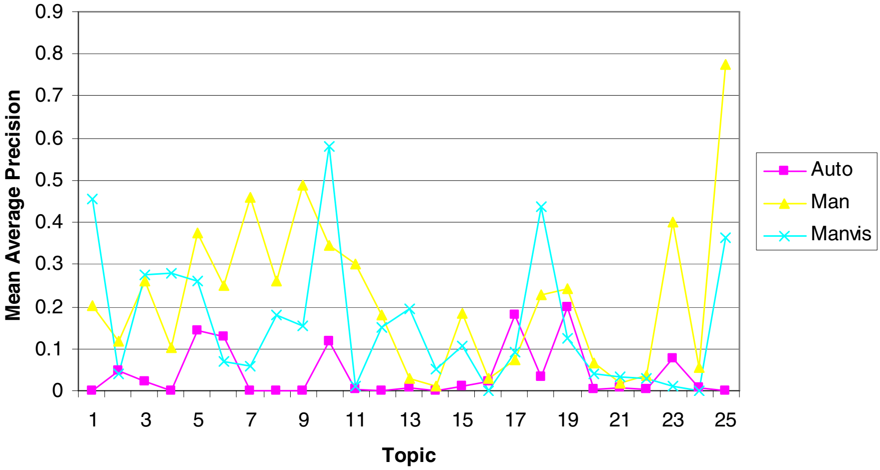

Fig. 3. Mean average precision for each of the three runs for each topic

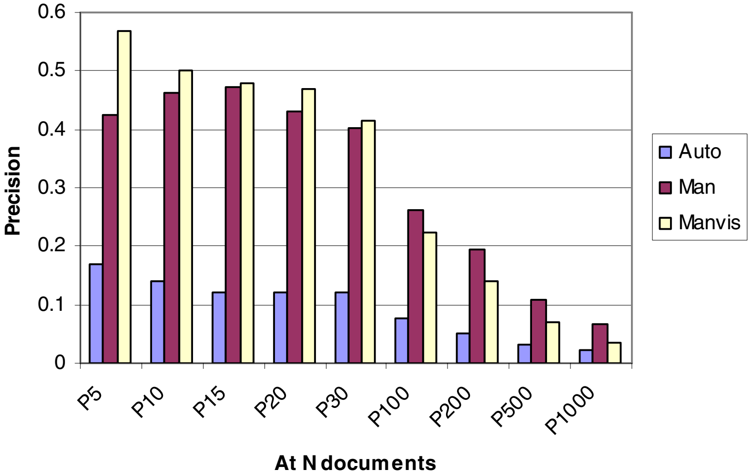

Fig. 4. Average of precision at 5, 10, 15, 20, 30, 100, 200, 500, and 1000 images for each run. The manual plus visual query run has higher precision down to 30 images retrieved, despite its lower mean average precision.

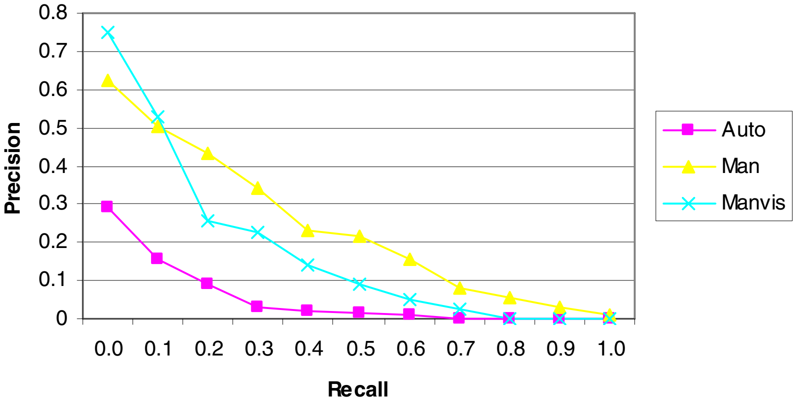

Fig. 5. Recall-precision curves for each run. The manual plus visual query run has a higher precision at low levels of recall (i.e., at the top of image output).

5 Conclusions

Our ImageCLEF medical track experiments showed that manual query modification and use of an automated translation tool provide benefit in retrieving relevant images. Filtering the output with findings from a baseline content-based approach diminished performance overall, but perhaps not in the part of the output most likely to be seen by real users, i.e., the top 30 images.

The experiments of our groups and others raise many questions about image retrieval:

- Which measures coming from automated retrieval evaluation experiments are most important for assessing systems in the hands of real users?
- How would text retrieval methods shown to be more effective in some domains (e.g., Okapi weighting) improve performance?
- How would better approaches to data fusion of semantic and visual queries impact performance?
- Are there methods of semantic and visual retrieval that improve performance in complementary manners?
- How much do these variations in output matter to real users?

Our future work also includes building a more robust image retrieval system proper, which will both simplify further experiments as well as give us the capability to employ real users in them. With such a system, users will be able to manually modify queries and/or provide translation. Additional work we are carrying out includes better elucidating the needs of those who use image retrieval systems based on a pilot study we have performed [6].

References

1. Clough, P., Müller, H., Deselaers, T., Grubinger, M., Lehmann, T., Jensen, J., Hersh W.: The CLEF 2005 cross-language image retrieval track. In: Springer Lecture Notes in Computer Science (LNCS), Vienna, Austria. (2006 - to appear)
2. Müller, H., Michoux, N., Bandon, D., Geissbuhler, A.: A review of content-based image retrieval systems in medical applications-clinical benefits and future directions. International Journal of Medical Informatics, **73** (2004) 1-23
3. Antani, S., Long, L., Thoma, G.R.: A biomedical information system for combined content-based retrieval of spine x-ray images and associated text information. Proceedings of the 3rd Indian Conference on Computer Vision, Graphics and Image Processing (ICVGIP 2002), Ahamdabad, India (2002)
4. Belkin, N., Cool, C., Croft, W.B., Callan, J.P.: Effect of multiple query representations on information retrieval system performance. In: Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Pittsburgh, PA. ACM Press (1993) 339-346
5. Cohen, A.M., Bhupatiraju, R.T., Hersh, W.: Feature generation, feature selection, classifiers, and conceptual drift for biomedical document triage, In: The Thirteenth Text Retrieval Conference: TREC 2004 (2004) http://trec.nist.gov/pubs/trec13/papers/ohsu-hersh.geo.pdf
6. Hersh, W., Jensen, J., Müller, H., Gorman, P., Ruch, P.: A qualitative task analysis of biomedical image use and retrieval. MUSCLE/ImageCLEF Workshop on Image and Video Retrieval Evaluation, Vienna, Austria (2005) http://medir.ohsu.edu/~hersh/muscle-05-image.pdf