Manual Query Modification and Automated Translation to Improve Cross-Language Medical Image Retrieval

Jeffery R. Jensen
William R. Hersh
Department of Medical Informatics & Clinical Epidemiology
Oregon Health & Science University
Portland, OR, 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 developed a context-based retrieval system. We then conducted our experiment using an automatic query, a manual query, and a manual/visual query. The best results were obtained from manual modification of queries, while combining those results with data from content-based image retrieval had a detrimental effect.

Categories and Subject Descriptors

H.3 [Information Storage and Retrieval]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; H.3.4 Systems and Software; H.3.7 Digital Libraries

General Terms

Measurement, Performance, Experimentation

Keywords

Question answering, Questions beyond factoids

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 40,000 images - annotated in English, French, and/or German - and 25 topics provided the basis for experiments.

There are two general approaches to image retrieval, context-based and content-based (Müller, Michoux et al., 2004). Context-based (or semantic) image retrieval uses a textual source of information to determine an image's subject matter, such as an annotation or more structured metadata. Content-based (or visual) image retrieval, on the other hand, uses features from the image, such as color, texture, shapes, etc., to determine its subject matter. The latter has historically been a difficult task, especially in the medical domain (Antani, Long et al., 2002). The most success has been for "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 skin cancers." Some researchers have begun to investigate hybrid methods that combine both image context and content for indexing and retrieval (Antani, Long et al., 2002).

Oregon Health & Science University (OHSU) participated in the medical image retrieval task of ImageCLEF. Our experiments were based on a context-based image retrieval system we developed, although we also attempted to improve our performance by augmenting the image output with results made available from a content-based search. Our experimental runs included an automatic query, a manually

modified query with automated translation, and a manual/visual query (the manual query refined with the results of a content-based search).

2 System Overview

Our web-based system was developed using open-source products from the Apache Software Foundation, such as Lucene and Tomcat. The core of the system is Java Servlet that uses Lucene and other tools to index and retrieve the text of annotations ("documents") associated with each image. The Servlet uses two main components, a Library and Lucene. The Library builds a data structure to manage the ImageCLEF data, whereas Lucene provides the functionality for the Indexer and Searcher objects. The Indexer indexes the textual annotations of the images, while the Searcher retrieves documents from the indexed data. Tomcat is used to host the system and administer basic levels of security for the ImageCLEF data. Figure 1 shows a graphical overview of our system.

The system works as follows. The query is sent to the system as a search-request. The Servlet then forwards the search-request to the Searcher object. The Searcher object performs the search and builds a result set from the Library. The result set is then returned to the Servlet to be displayed to the user. The user can then navigate through the results or perform a new search, if desired.

In order to index the image annotations, we developed a library component to represent the data. The library consists of collections (CASImage, MIR, PathoPic and Peir), cases, images and their respective annotations. Each collection consists of one or more cases; each case contains one or more images, and one or more annotations; and each image contains one or more annotations. Each annotation is an XML document describing either the case or an image.

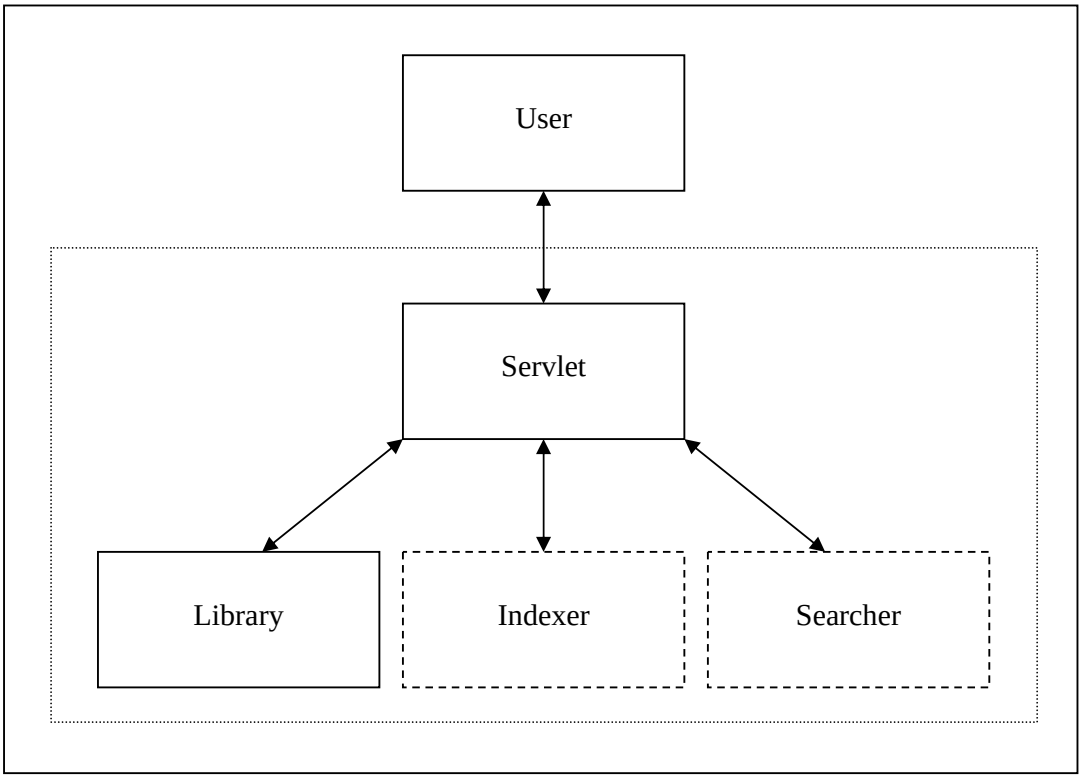

Figure 1 - The dashed line represents the components that were developed from Lucene, and Tomcat, which is represented by the dotted line, hosts the entire system.

Apache Lucene is a freely available and fully featured search engine (lucene.apache.org), which we have used in other retrieval evaluation forums, such as the Text Retrieval Conference (TREC) Genomics Track (Cohen, Hersh et al., 2005). Documents are indexed by parsing of individual words within them and weighting via inverse document frequency (IDF) and term frequency (TF) functions.

Lucene is distributed with a variety of Analyzers for textual indexing. We chose Lucene's Standard Analyzer, which supports acronyms, floating point numbers, as well as the 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. However, we indexed each annotation without the use of an XML parser. Therefore, every element was indexed along with its corresponding value.

Lucene provides an application program interface (API) and robust query language that supports a variety of term modifiers and searching operators. It supports two different types of terms for searching, single and phrase. These terms can be searched over the entire Library, or over specific data fields. As our data contains specific fields (the collection name, file name, and the contents of the file to be indexed), users can search over a specific collection or retrieve a known case or image.

As with most information retrieval systems, Lucene uses a scoring algorithm that sums for each query term in each document the product of the TF, the 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

3 Retrieval

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

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

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. Therefore, the search statements for this run were 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 2 shows a sample query from our second run, OHSUman.

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

Figure 2 - Manual query for topic 12.

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.

Our third run (OHSUmanviz) used a combination of textual and visual searching methods. We took the image output from OHSUman and excluded all documents that were not retrieved by the University of Geneva content-based "baseline" 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" context-based and content-based run.

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 placed in the "text-only manual" category and achieved an MAP of 0.2116. Despite being the only submission in this category, this technique managed to score above 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. We were surprised to see that 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. Therefore the difference in the overall score was a result of the visual refinement.

Figure 3 illustrates the number of images returned by each of the techniques, while Figure 4 show MAP per query 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.

We also noticed that all of our techniques did not perform well for queries 11-17 and 20-22. The majority of these topics were seeking radiographs, ranging from x-rays, CT scans, and MRI scans. We hope to carry out further analysis to identify other patterns where our approaches failed and, more importantly, where other techniques may provide benefit.

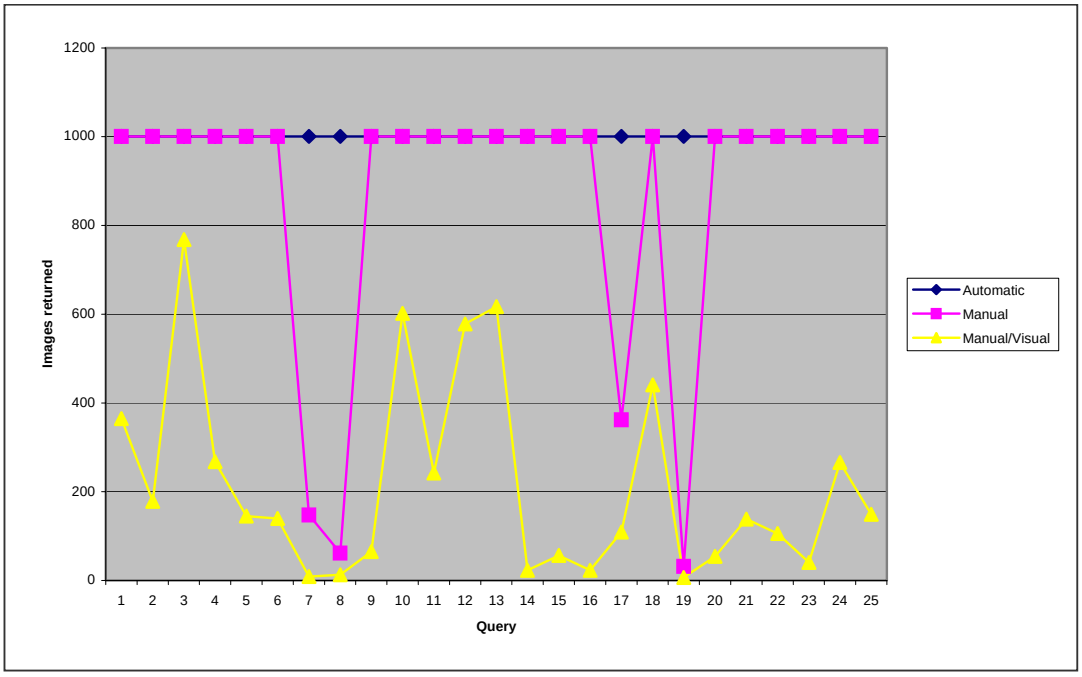

Figure 3 - Number of images returned per query, for each technique

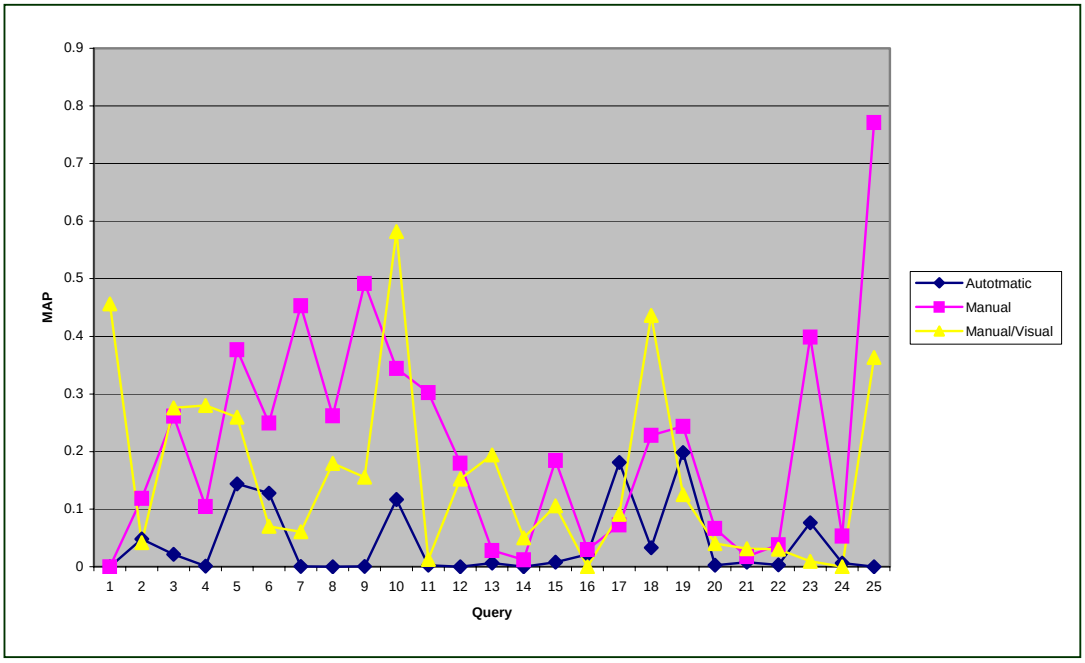

Figure 4 - Mean average precision (MAP) per query for each technique

5 Conclusions

Our ImageCLEF medical track experiments showed that manual query modification and use of an automated translation tool provide substantial benefit in retrieving relevant images. Filtering the output with findings from a baseline content-based approach diminished performance. There are likely additional permutations of these techniques that would improve performance, and future plans include testing them to identify which ones do so. We also hope to explore other methods, such as more selective use of the textual annotations, use of more advanced retrieval weighting algorithms (e.g., Okapi weighting), and better integration with content-based imaging systems.

We also aim to build a more robust image retrieval system proper, which will both simplify further experiments as well as give us the capability to employ real users to better manually modify queries and/or provide translation. Additional work we aim to carry out includes better elucidating the needs of those who use image retrieval systems to better inform our own system as well as test collections to assess systems of others and ourselves.

6 References

Antani, S., Long, L., et al. (2002). 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.

Cohen, A., Hersh, W., et al. (2005). Using co-occurrence network structure to extract synonymous gene and protein names from MEDLINE abstracts. *BMC Bioinformatics*, 6: 103. http://www.biomedcentral.com/1471-2105/6/103.

Müller, H., Michoux, N., et al. (2004). A review of content-based image retrieval systems in medical applications-clinical benefits and future directions. *International Journal of Medical Informatics*, 73: 1-23.