Improving Retrieval Using External Annotations: OHSU at ImageCLEF 2010

Steven Bedrick and Jayashree Kalpathy-Cramer

Oregon Health & Science University, Portland, OR, USA 97239

Abstract. Over the past several years, our team has focused its efforts on improving retrieval precision performance by mixing visual and textual information. This year, we chose to explore ways in which we could use external data to enrich our retrieval system's data set; specifically, we annotated each image in the test collection with a set of MeSH headings from two different sources: human-assigned MEDLINE index terms, and automatically-assigned MeSH headings (via the National Library of Medicine's MetaMap software).

In addition to exploring these different data enrichment techniques, we also revamped the architecture of our retrieval system itself. In past years, we have used a two-tiered approach wherein the data is stored in a relational database (RDBMS), but the indexing and searching are done using Lucene-like system. This year, we took advantage of our RDBMS's full-text search capabilities and performed both storage and searching in the RDBMS. This turned out to have both positive and negative effects at a practical level. On the one hand, using the database's built-in text retrieval subsystem resulted in improved retrieval speed and easier query analysis; however, these gains came at the cost of reduced flexibility and increased code complexity.

Our experiments investigated the effects of using various combinations of human- and automatically-assigned MeSH terms, along with several of the techniques that have proved useful in previous years. We found that including automatically-assigned MeSH terms sometimes provided a small amount of improvement (in terms of bpref, MAP, and early precision) and sometimes hurt performance, whereas including the human-assigned MEDLINE index headings consistently yielded a sizable improvement in those same metrics.

1 Introduction

As has been discussed at length in previous works[12, 11], medical image retrieval represents a large and ever-growing problem. As the utilization rates of diagnostic imaging increase[9, 4, 10, 3], so do the number of images that must be stored and retrieved. Unfortunately, however, image retrieval techniques often lag behind their textual cousins in terms of performance[6].

The ImageCLEF series of evaluation campaigns provides a forum for researchers working on image retrieval problems to share ideas and compare their systems. One of the campaign's ongoing tracks is a medical image retrieval task, which is described in detail in [12, 11]. This year, the task's test collection was an expanded version of the collection used in 2008 and 2009, and included 77,495 images from 5,609 articles in the journals Radiology and Radiographics.

OHSU has participated in the medical track of ImageCLEF since 2006, and our focus has consistently been on exploring ways to make use of both visual and textual information during retrieval. Over the past several years, our system has achieved good performance (particularly in terms of precision) by using image modality information to adaptively determine which results are relevant to a given query[5, 8, 13, 7].

This year, however, we decided to try something slightly different. In the past, we had annotated documents in the test collection with automatically-extracted modality labels and made use of external knowledge sources (such as the US National Library of Medicine's UMLS metathesaurus) to perform query expansion. This year, we attempted to make further use of external knowledge in the form of MeSH (Medical Subject Heading) keyword annotations.

Each record in the ImageCLEF medical test collection includes a "PMID," or PubMed identifier: essentially, a link back to the MEDLINE record (journal article) from which the image originally came. Each entry in MEDLINE is indexed by a professional indexer, and we added these index terms to our copy of the test collection. We therefore were able to make use of an average of 12 guaranteed-relevant keywords for each item in the collection. Furthermore, since a certain number of index terms for each MEDLINE entry are designated as "major headings" (i.e., particularly relevant keywords), we were able to be highly confident of the relevance of at least a few index terms for each collection entry.

Of course, most medical data sets do *not* include human-curated index terms. We therefore experimented with automatically-assigned index terms using the National Library of Medicine's MetaMap software[1][1]. MetaMap identifies concepts in arbitrary input text, and maps them to UMLS concepts. We used MetaMap to assign a set of MeSH headings to each image caption in the test collection[2]. MetaMap assigned an average of 5 MeSH terms to each caption.

2 Our System

As in years past, our retrieval system is written in the Ruby language[3] and uses the Ruby on Rails[4] web framework as well as the open-source PostgreSQL relational database system[5]. However, unlike our system from 2006–2009, this year's system uses neither Lucene nor Ferret (a port of Lucene to Ruby)[6] to perform

[1] `http://mmtx.nlm.nih.gov/`

[2] See [2] for an up-to-date discussion on the current state of MetaMap.

[3] `http://www.ruby-lang.org`

[4] `http://www.rubyonrails.org`

[5] `http://www.postgresql.org`

[6] To minimize confusion, we will refer our past systems as having used Lucene, as in terms of capabilities, APIs, and query language, Ferret is functionally identical to Lucene.

the text retrieval. Instead, this year we chose to experiment with PostgreSQL's built-in full-text search subsystem[7].

In the past, our system had to maintain its full-text index of image captions, titles, etc. as a separate file, and relied on extra software libraries to perform retrieval. Our hope was that, by integrating the full-text searching with the database itself, our system would have fewer "moving parts." This turned out to be the case; integrating text search with the database did make certain parts of our system less cluttered, and this year's system's retrieval speed was definitely improved over previous years' systems.

However, these improvements came at a cost. PostgreSQL's full-text query syntax is somewhat cumbersome, and modifying our query processor to translate user queries into PostgreSQL-compatible queries turned out to be more challenging than we had initially anticipated. That said, the fact that PostgreSQL's search subsystem uses a set of extensions to standard SQL syntax meant that it was extremely convenient and easy to develop and debug our query processor, particularly in comparison to our analogous experiences with Lucene, whose query system could be somewhat opaque.

Overall, though, the final product ended up being somewhat more complex than its Lucene-based predecessor, and there were also certain convenience features that we missed. For example, the port of Lucene that we used made it very easy to add additional index fields that represented dynamically calculated values (for example, a concatenation of two other fields). Achieving a similar effect using PostgreSQL involved adding a new index to our database itself. While this is certainly easy enough to do, it ultimately resulted in a very cluttered database schema.

Overall, integrating our search system with the database itself was probably a wash from a technical standpoint. From a performance standpoint, it ended up representing a small step backwards in certain ways. Lucene uses a vector-space retrieval model, whereas PostgreSQL's text search subsystem uses a boolean model. As such, we found this year's system to be much less robust when faced with topics with few relevant results, which affected its recall.

Besides this architectural change, other aspects of our system were relatively unchanged from the descriptions given in previous years' Working Notes papers[13, 7], including modality filtration, query expansion, etc. One extension that we did add over previous years was "modality-aware result reordering." In past years, we had found that our existing modality filtration techniques[8] were sometimes too aggressive, particularly in situations where the modality information was ambiguous or where there were not very many relevant results in the collection.

To compensate for this, we added a mode to our system wherein the final result set returned to the user will contain both filtered and un-filtered results,

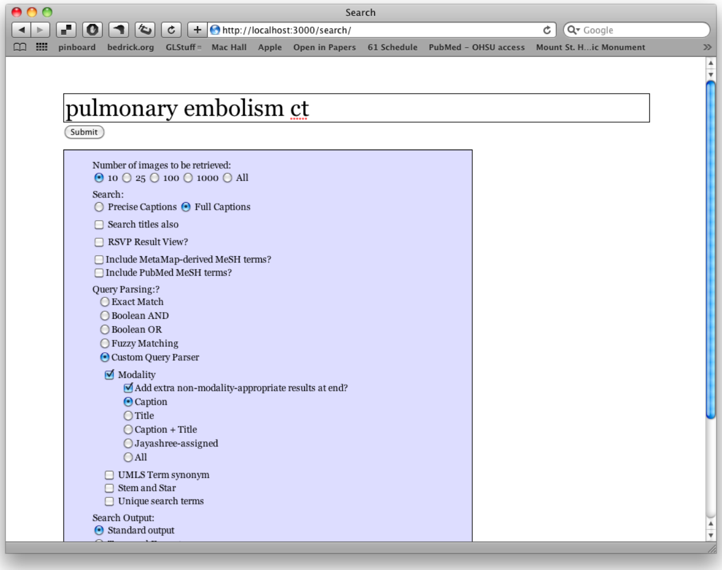
[7] `http://www.postgresql.org/docs/8.4/static/textsearch.html`

[8] See [7, 13] for more details; in brief, our modality filtration approach involves extracting modality information from user queries, and then returning only result images whose modality matches that specified by the user.

but with the filtered results ordered above the unfiltered ones. In other words, if a user searches for “Pulmonary embolism CT”, the system’s results will be composed of two subsets: first, the results whose modalities were CT scans, and then, any other search results that might have come up with other modalities. Our intent was to improve recall performance over simple modality filtration runs.

In addition to this modification, as mentioned earlier, this year we also added externally-derived annotations to our database, in the form of MeSH headings. To actually make use of these for retrieval, we combined the image caption fields with a space-delimted list of MeSH headings to create two “meta-fields” (one that included the human-generated MEDLINE index terms, and another for the automatically-assigned headings). We then created full-text indices of these columns just as we would any others (e.g., caption, title, etc.), and used them accordingly.

As in previous years, our system can be used interactively (see figures 1 and 2). However, it can also be used in a more batch-oriented manner (see figure 3), and its results can be downloaded directly as `trec-eval`-compliant run files. Furthermore, one of the authors (SB) has written a script which submits queries and generates runs programmatically.



pulmonary embolism ct

Number of images to be retrieved:
● 10 ○ 25 ○ 100 ○ 1000 ○ All

Search:
○ Precise Captions ● Full Captions

□ Search titles also

□ RSVP Result View?

□ Include MetaMap-derived MeSH terms?
□ Include PubMed MeSH terms?

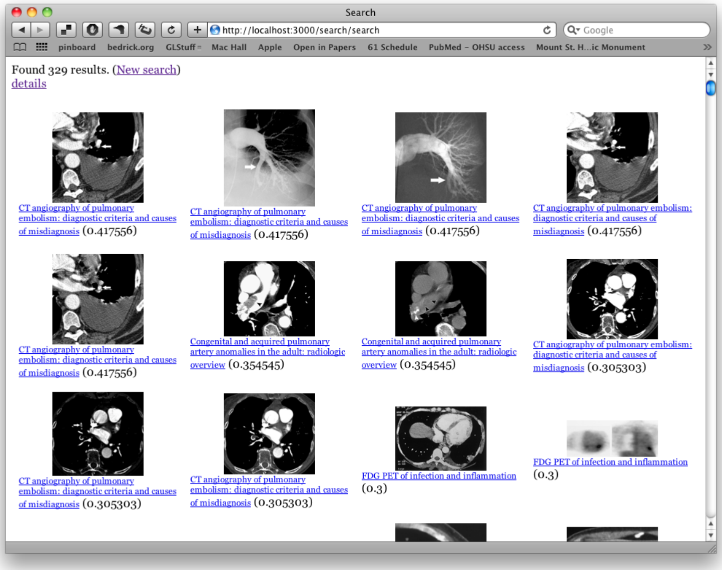
Query Parsing:?
○ Exact Match
○ Boolean AND
○ Boolean OR
○ Fuzzy Matching
● Custom Query Parser

☑ Modality
☑ Add extra non-modality-appropriate results at end?
● Caption
○ Title
○ Caption + Title
○ Jayashree-assigned
○ All

□ UMLS Term synonym
□ Stem and Star
□ Unique search terms

Search Output:
● Standard output

Fig. 1. Our system’s main search screen. Note the large number of options available for controlling search behavior.



Found 329 results. (New search)
details

CT angiography of pulmonary embolism: diagnostic criteria and causes of misdiagnosis (0.417556)

CT angiography of pulmonary embolism: diagnostic criteria and causes of misdiagnosis (0.417556)

CT angiography of pulmonary embolism: diagnostic criteria and causes of misdiagnosis (0.417556)

CT angiography of pulmonary embolism: diagnostic criteria and causes of misdiagnosis (0.417556)

CT angiography of pulmonary embolism: diagnostic criteria and causes of misdiagnosis (0.417556)

Congenital and acquired pulmonary artery anomalies in the adult: radiologic overview (0.354545)

Congenital and acquired pulmonary artery anomalies in the adult: radiologic overview (0.354545)

CT angiography of pulmonary embolism: diagnostic criteria and causes of misdiagnosis (0.305303)

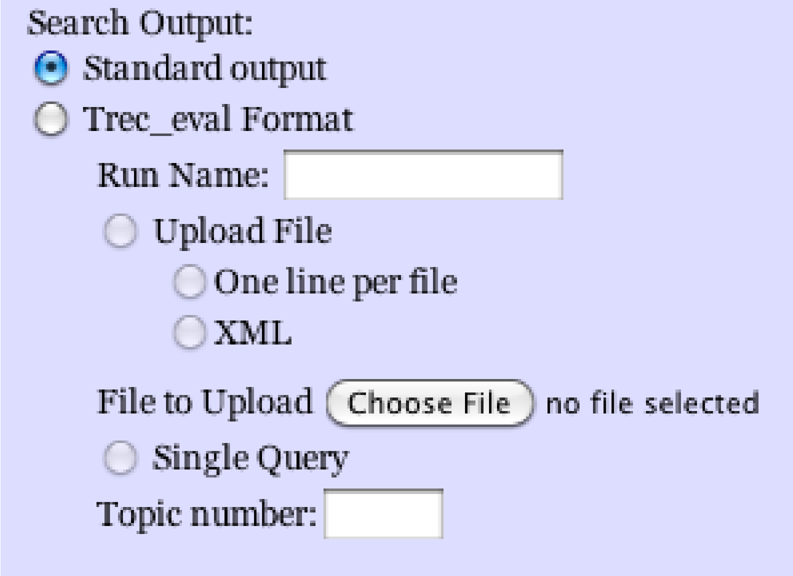
CT angiography of pulmonary embolism: diagnostic criteria and causes of misdiagnosis (0.305303)

CT angiography of pulmonary embolism: diagnostic criteria and causes of misdiagnosis (0.305303)

FDG PET of infection and inflammation (0.3)

FDG PET of infection and inflammation (0.3)

Fig. 2. The system's main search results screen. Clicking on a single result takes the user to a full page with details of that specific image.



Search Output:

● Standard output

○ Trec_eval Format

Run Name:

○ Upload File

○ One line per file

○ XML

File to Upload no file selected

○ Single Query

Topic number:

Fig. 3. The system can be configured to give results in an interactive, user-friendly screen (see figure 2), or it can be set to return a trec_eval-compatible file as its output. Queries can be submitted one at a time, or a collection of queries may be uploaded in a variety of formats.

bar	run name	MetaMap	MEDLINE	Modality	Reorder	Expansion	Titles
a	pm_all_all_mod	no	all	all	no	yes	no
b	pm_major_all_mod	no	major	all	no	yes	no
c	all_mh_major_all_mod	yes	major	all	no	yes	no
d	all_mh_major_jaykc_mod	yes	major	jaykc	no	yes	no
e	high_recall_with_titles_modality_reorder	yes	all	all	yes	yes	yes
f	high_recall_with_titles	yes	all	no	no	yes	yes
g	all_mh_major_jaykc_mod_reorder	yes	major	jaykc	yes	yes	no
h	all_mh_major_all_mod_reorder	yes	major	all	yes	yes	no
i	mm_all_mod	yes	no	all	no	yes	no
j	high_recall	yes	all	no	no	yes	no
k	*control*	*no*	*no*	*jaykc*	*yes*	*yes*	*yes*

Table 1. Key for figure 4. Runs are ordered by MAP, except for control. MetaMap: whether MetaMap-derived MeSH headings were used; MEDLINE: whether MEDLINE index headings were used, and, if so, whether only "major subject" headings were used; Modality: whether modality filtration was used, and, if so, which classifier's modalities were used; Reorder: whether modality reordering was used (see 2 for details); Expansion: whether UMLS query expansion was performed; Titles: whether image titles were used in addition to captions for retrieval.

3 Runs Submitted

We submitted a total of ten ad-hoc runs, all of which used some combination of system features (see table 1 for a breakdown of the runs). Our primary focus this year was on exploring the effects of using different combinations of MeSH terms. As mentioned earlier, we had two types of MeSH term for each record in the collection: a set of human-assigned MEDLINE index terms, and a set of automatically-assigned terms. Of the MEDLINE terms, some were designated by the human indexers at the National Library of Medicine as "major topics" (i.e., particularly relevant key words). Our runs either did or did not include the automatically-assigned MeSH terms ("MetaMap"), and used either none of the MEDLINE terms, only the major topic terms, or all of the MEDLINE terms.

Another setting we varied from run to run was whether to use modality filtration, and, if so, whether to use modalities extracted from image titles, captions, a visual classifier ("jaykc"), or the union of all three. Additionally, some of our runs used image titles as well as captions; others only used captions.

For the purpose of comparison, we also produced a "control" run featuring all of our "classic" features (query expansion, modality filtration and reordering, and use of titles) but without any of the MeSH terms. We did not submit this run to ImageCLEF, but we did run it through `trec_eval` with the same qrel file. As such, its results are directly comparable to those of our submitted runs.

4 Results and Discussion

OHSU's runs performed competitively, although we were not at the very top of the categories we competed in. Our results are summarized in table 2 and figure 4. In terms of both MAP and bpref, our runs follow a bimodal distribution, with

run	bpref	map	p5	p10	p20	p100
pm_all_all_mod	0.344	0.3029	0.4875	0.4313	0.3344	0.1562
pm_major_all_mod	0.3404	0.3004	0.5000	0.4375	0.3469	0.1519
all_mh_major_all_mod	0.3428	0.2983	0.4625	0.4188	0.3031	0.1494
all_mh_major_jaykc_mod	0.3428	0.2983	0.4625	0.4188	0.3031	0.1494
high_recall_with_titles_modality_reorder	0.2754	0.2623	0.4375	0.3875	0.293	0.1644
high_recall_with_titles	0.2714	0.2592	0.4375	0.3875	0.2937	0.1581
all_mh_major_jaykc_mod_reorder	0.2533	0.256	0.4375	0.3813	0.275	0.1487
all_mh_major_all_mod_reorder	0.2533	0.256	0.4375	0.3813	0.275	0.1487
mm_all_mod	0.2594	0.2476	0.4625	0.4125	0.3062	0.1444
high_recall	0.2533	0.2386	0.4125	0.3625	0.2844	0.1544
control	*0.2614*	*0.2397*	*0.4000*	*0.3625*	*0.2875*	*0.1581*

Table 2. OHSU runs submitted for ImageCLEF 2010 (sorted by MAP, except for control), along with a "control" run using none of the external MeSH annotations described in section 2.

runs a–d noticeably outperforming the remainder. In terms of early precision, our runs all performed quite well, although there was some notable between-run variation.

Regarding the performance of runs a–d, the question arises: what was different about those runs from the others? They all featured the combination of MEDLINE terms, modality filtration, query expansion, and *no* result reordering. Similarly-configured runs that *did* use result reordering suffered a map penalty, as did runs without reordering but using MetaMap terms instead of MEDLINE terms (e.g., `mm_all_mod`, although it should be noted that while this particular run had a low map, it did quite well in terms of simple early precision). Using only "major subject" MEDLINE headings did not seem to yield any significant benefit over simply using all of the human-assigned index terms.

5 Conclusions

In conclusion, our idea of using external MeSH annotations to improve retrieval does seem to have promise; however, there is a great deal of experimentation yet to do before we can take full advantage of these annotations. Specifically, we need to determine why the manually-assigned MEDLINE annotations seemed to be so much more beneficial than the automatically-assigned annotations. One possible explanation would be that the average quality of the automatically-assigned annotations is lower than the average quality of the human-assigned annotations— i.e., that the MetaMap program is assigning erroneous or incomplete subject headings. Our initial examinations have actually shown the opposite to be true— given image captions, MetaMap seems to be doing a reasonable job at assigning relevant and specific MeSH headings.

Another possibility is that the level of annotation detail is different between the human- and automatically-assigned MeSH headings, and that this differ-

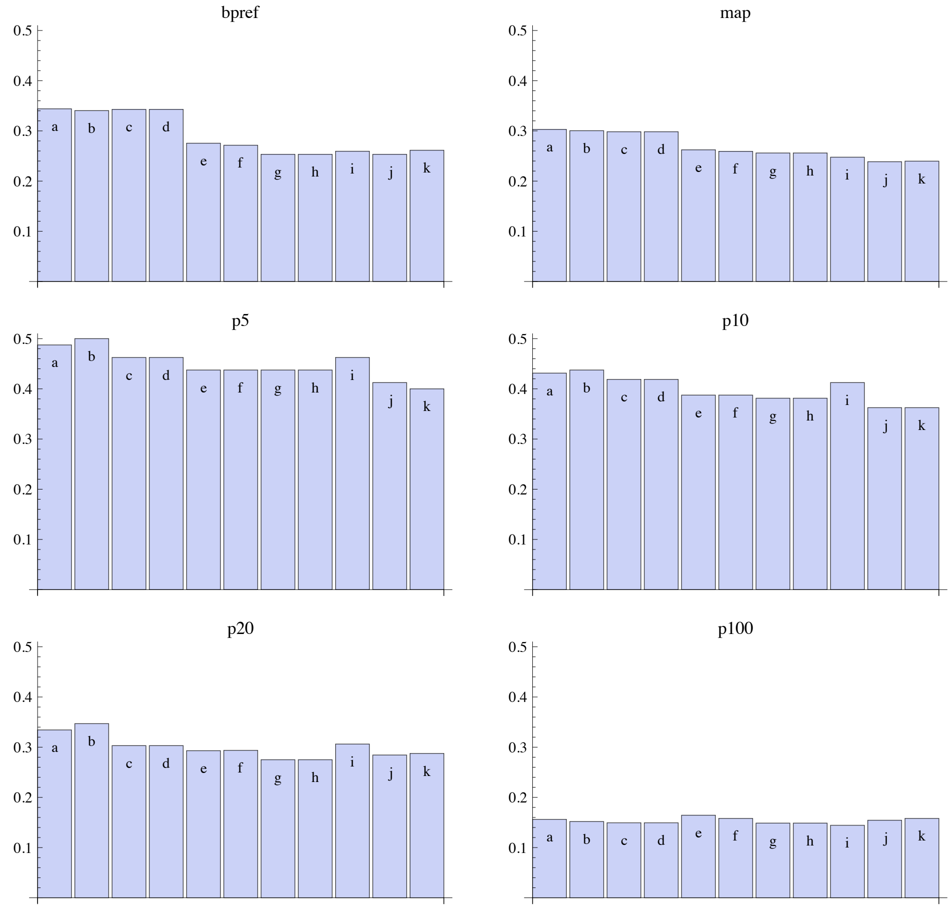


Fig. 4. Results for the ten runs submitted, along with the "control" run as described in section 3. See table 1 for the key.

ence in detail is affecting the ability of the terms to help retrieval. The human indexers are assigning subject headings for an entire article, whereas in our system MetaMap is working on individual captions. Therefore, our automatically-assigned annotations sometimes seem to be very specific, whereas the human-assigned annotations can seem very vague. For example, consider the case of figure 62141, from an article entitled "High-resolution CT and CT angiography of peripheral pulmonary vascular disorders." The caption text is as follows:

> Figure 8c. Parasitic pulmonary embolism. (a, b). CT scans demonstrate rupture of an *Echinococcus* cyst (*E. granulosus*) (*) into the inferior vena cava (c, d). CT scans show peripheral pulmonary embolism of scolices (c) with subpleural calcified daughter cysts (arrowheads in d). Massive central pulmonary arterial embolism can occur in hydatid disease or in ascariasis in association with acute pulmonary arterial thrombosis.

This represents a detailed description of a figure, including lots of helpful anatomical vocabulary. The MEDLINE terms for this article are:

- Angiography
- Humans
- Lung Diseases
- Peripheral Vascular Diseases
- Tomography, X-Ray Computed

The MetaMap-derived terms, on the other hand, are as follows:

- Rupture
- Cysts
- Thrombosis
- Echinococcosis
- Ascariasis
- Pulmonary Artery
- Tomography, X-Ray Computed
- Vena Cava, Inferior
- Pulmonary Embolism

Clearly, these two sets of terms are operating at different levels of specificity. While we have not done an exhaustive study of our annotations, this pattern does seem to repeat itself with some regularity from record to record. Given this asymmetry in annotation detail between sources, it is not surprising that one source would prove more helpful than the other. However, the questions of *which* source would be most helpful, and, more importantly, *why* it would be so, remain open and will form the next step in this research.

Additional future steps include exploring ways to enrich our use of these annotations. Currently, we are using them in a very simplistic way, and are treating all annotations as equally important. Perhaps we should only make use of certain categories of annotation, and use only anatomical terms, for example.

We may want to weight certain annotations more heavily than others, based perhaps on frequency of occurrence. Some very common annotations might be best ignored altogether. Hopefully, exploring these directions will enable us to improve our system's performance for next year's ImageCLEF.

Acknowledgements

We acknowledge the National Library of Medicine grant 3T15LM007088 for supporting this work, as well as 1K99LM009889.

References

1. Aronson, A.R.: Effective mapping of biomedical text to the umls metathesaurus: the metamap program. Proc AMIA Symp pp. 17–21 (2001)
2. Aronson, A.R., Lang, F.M.: An overview of metamap: historical perspective and recent advances. J Am Med Inform Assoc 17(3), 229–36 (May 2010)
3. Bhargavan, M., Sunshine, J.H.: Utilization of radiology services in the united states: levels and trends in modalities, regions, and populations. Radiology 234(3), 824–32 (Mar 2005)
4. Dinan, M.A., Curtis, L.H., Hammill, B.G., Patz, Jr, E.F., Abernethy, A.P., Shea, A.M., Schulman, K.A.: Changes in the use and costs of diagnostic imaging among medicare beneficiaries with cancer, 1999-2006. JAMA 303(16), 1625–31 (Apr 2010)
5. Hersh, W., Kalpathy-Cramer, J., Jensen, J.: Medical image retrieval and automated annotation: Ohsu at imageclef 2006. In: Peters, C., Clough, P., Gey, F.C., Karlgren, J., Magnini, B., Oard, D.W., de Rijke, M., Stempfhuber, M. (eds.) CLEF. Lecture Notes in Computer Science, vol. 4730, pp. 660–669. Springer (2006), http://dblp.uni-trier.de/db/conf/clef/clef2006.html#HershKJ06
6. Hersh, W.R., Müller, H., Jensen, J.R., Yang, J., Gorman, P.N., Ruch, P.: Advancing biomedical image retrieval: Development and analysis of a test collection. J Am Med Inform Assoc p. M2082 (Jun 2006), http://www.jamia.org/cgi/content/abstract/M2082v1
7. Kalpathy-Cramer, J., Bedrick, S., Hatt, W., Hersh, W.: Multimodal medical image retrieval ohsu at imageclef 2008. In: Peters, C., Deselaers, T., Ferro, N., Gonzalo, J., Jones, G., Kurimo, M., Mandl, T., Peñas, A., Petras, V. (eds.) Evaluating Systems for Multilingual and Multimodal Information Access, pp. 744–751. Lecture Notes in Computer Science, Springer Berlin / Heidelberg, 10.1007/978-3-642-04447-2_96
8. Kalpathy-Cramer, J., Hersh, W.: Medical image retrieval and automatic annotation: Ohsu at imageclef 2007. In: Peters, C., Valentin, J., Mandl, T., Müller, H., Oard, D., Peñas, A., Petras, V., Santos, D. (eds.) Advances in Multilingual and Multimodal Information Retrieval: 8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007, Budapest, Hungary, September 19-21, 2007, Revised Selected Papers. Lecture Notes in Computer Science, vol. 5152, pp. 623–630. Springer-Verlag, Berlin, Heidelberg (2008)
9. Maitino, A.J., Levin, D.C., Parker, L., Rao, V.M., Sunshine, J.H.: Nationwide trends in rates of utilization of noninvasive diagnostic imaging among the medicare population between 1993 and 1999. Radiology 227(1), 113–7 (Apr 2003)
10. Maitino, A.J., Levin, D.C., Parker, L., Rao, V.M., Sunshine, J.H.: Nationwide trends in rates of utilization of noninvasive diagnostic imaging among the medicare population between 1993 and 1999. Radiology 227(1), 113–7 (Apr 2003)

11. Müller, H., Kalpathy-Cramer, J., Eggel, I., Bedrick, S., Radhouani, S., Bakke, B., Jr., C.K., Hersh, W.: Overview of the medical retrieval task at imageclef 2009. Working Notes of the CLEF 2009 workshop, Corfu, Greece (2009)
12. Müller, H., Kalpathy-Cramer, J., Kahn Jr., C.E., Hatt, W., Bedrick, S., Hersh, W.: Overview of the ImageCLEFmed 2008 medical image retrieval task. In: Peters, C., Giampiccol, D., Ferro, N., Petras, V., Gonzalo, J., Peñas, A., Deselaers, T., Mandl, T., Jones, G., Kurimo, M. (eds.) Evaluating Systems for Multilingual and Multimodal Information Access – 9th Workshop of the Cross-Language Evaluation Forum. Lecture Notes in Computer Science, Aarhus, Denmark (Sep 2008 (printed in 2009))
13. Radhouni, S., Kalpathy-Cramer, J., Bedrick, S., Bakke, B., Hersh, W.: Multimodal medical image retrieval improving precision at imageclef 2009. In: Peters, C. (ed.) Working Notes for the CLEF 2009 Workshop, Corfu, Greece (2009)