MEDINFO 2007
K. Kuhn et al. (Eds)
IOS Press, 2007

Analyzing Web Log Files of the Health On the Net HONmedia Search Engine to Define Typical Image Search Tasks for Image Retrieval Evaluation

Henning Müller [a], Célia Boyer [b], Arnaud Gaudinat [b], William Hersh [c], Antoine Geissbuhler [a]

[a] *Medical Informatics Service, University and Hospitals of Geneva, Geneva, Switzerland*
[b] *Health On the Net (HON), Geneva, Switzerland*
[c] *Oregon Health and Science University (OHSU), Portland, OR, USA*

Abstract

Medical institutions produce ever-increasing amount of diverse information. The digital form makes these data available for the use on more than a single patient. Images are no exception to this. However, less is known about how medical professionals search for visual medical information and how they want to use it outside of the context of a single patient. This article analyzes ten months of usage log files of the Health on the Net (HON) medical media search engine. Key words were extracted from all queries and the most frequent terms and subjects were identified. The dataset required much pre-treatment. Problems included national character sets, spelling errors and the use of terms in several languages.

The results show that media search, particularly for images, was frequently used. The most common queries were for general concepts (e.g., heart, lung). To define realistic information needs for the ImageCLEFmed challenge evaluation (Cross Language Evaluation Forum medical image retrieval), we used frequent queries that were still specific enough to at least cover two of the three axes on modality, anatomic region, and pathology. Several research groups evaluated their image retrieval algorithms based on these defined topics.

Keywords:

log files analysis, image retrieval evaluation.

Introduction

An increasing amount of medical information is being produced digitally, making it available for further processing and use, i.e., for teaching and research. Much of the produced data and experiences from past cases can be used to create tools for diagnostic decision aid. A great deal of medical information is also available on the Internet, as there are increasing requests for medical information by patients and professionals [1]. MedLinePlus is one example of a repository created to inform non-professionals, patients searching for information. Another example is Health On the Net (HON), which develops quality criteria for medical web pages and has an accreditation service for pages adhering to several quality criteria. HON also runs web search engines for medical web content aimed at patients and medical professionals with a multilingual search interface[1] [2]. Much research has been done on the searching of medical texts [3] but less on how images are used and searched for, although the amount of image data being produced is rising [4]. Many medical image databases are available within institutions, mainly for teaching, but some are also made available on the Internet. These include Casimage, HEAL (Health Education Assets Library), MedPix, and the Pathopic datasets. MIRC[2] (Medical Image Resource Center) is an initiative of the Radiological Society of North America (RSNA) to unite teaching files under a single interface. These databases contain thousands of annotated images. Unfortunately, the images are only rarely indexed in search engines such as Google as they are usually only available through the search in database fields. Another problem is that the annotation is often incomplete and information on the image modality is not always given. A search for “lung CT” with Google image search in October 2005 brought 160 results, about half of them lung CTs. The abovementioned databases alone contain several thousand lung CTs.

Outside of medicine, visual information retrieval has been an extremely active research domain for more than 15 years [5]. Studies on domain-specific user requirements have been performed, for example for journalists searching for images [6] or in the cultural heritage domain [7]. In the medical field, visual information retrieval has been proposed many times as extremely useful [8, 9]. Still, most research has a limited focus on retrieval for one particular group of images [9]. Although this might be a domain with high potential impact, teaching and research are more likely to profit first from possibilities to browse very large and diverse image collections by visual properties. In the context of ImageCLEFmed [10], a challenge evaluation for medical image retrieval, two surveys were performed among medical image users [11, 12] to find out more about typical information needs and search tasks. CLEF (Cross-Language Evaluation Forum) is a challenge evaluation for retrieval of multilingual information. ImageCLEFmed in particular evaluates the quality of retrieval from multilingual medical image retrieval available on the Internet. The

[1] http://www.wrapin.org/ & http://www.hon.ch/HONselect/
[2] http://mirc.rsna.org/

Selected for best paper award.

surveys include five user groups: medical professionals for diagnosis, teaching, and research as well as medical students and librarians. The goal of the work descried in this paper was to create realistic search tasks for ImageCLEFmed[3] based on information needs of web users. The analysis resulted in 30 search tasks used by participating research groups. Among the techniques used was analysis of log files, an active research domain [13], mainly to analyze web page design.

Materials and methods

Used data sets

The data used for this study were log files containing query terms of the HONmedia[4] search engine. The examined period of queries included ten months, from January 1, 2005 to October 31, 2005. This period was sufficient for a representative evaluation of search terms. Variations of search frequency or quality over the months were not part of our analysis. The original data set contained 53’970 queries. With each automatically extracted query term, the date and time of the query was stored. It was also stored whether the query was directly done via the HONmedia interface or referred to from Google towards HONmedia search. Many queries were in French, as the French-speaking medical community frequently uses the HON query engine. It was not possible to perform an automatic translation of the topics, as language detection is hard with only very few words. Other languages identified for the queries were English, German, Spanish, and Italian.

Pre-treatment of the data and evaluation techniques

The analysis of the data was done on a Linux computer using Perl to analyze the text files. The original data sets were transferred to pure text and the information on time and date of the query were discarded. Perl was used mainly to pre-treat the data. As data were extracted automatically and as robots perform queries on web interfaces there are many different formats for queries (sometimes broken), plus a variety of international character containing umlauts and accented characters sets that need to be combined.

Results

The data contained two groups of queries, queries directly asked via HON and queries forwarded via Google. These groups were treated separately. A total of 37’293 queries were directly performed via HONmedia and 16’677 were forwarded via Google.

Text normalization

First, normalization was necessary for the text to remove differences in coding of the strings, parameter options transmitted and for broken queries containing graphical symbols. We did not treat the word order in the queries. The steps were mainly based on a manual analysis of the data:

- Unify coding issues, to remove accents, Umlauts, national symbols, and any sort of non-text: –“()+–.
- Remove commands and options send by web robots or search engines.
- Remove URLs or fragments of URLs.
- Convert all characters to lower case.
- Change plural of frequent terms such as “images”.
- Remove frequent terms to define the target media: image(s) (5’796), media (512), video(s) (334).

Over 100 rules for normalization and removal were defined and applied to clean the data. Even after the removal steps, it was apparent that an extremely large number of different queries remained. In total, there remained 5’365 different unique queries (of 16’677) for the Google queries and 17’643 different HON queries (of 37’293). This meant that almost half the queries were unique being asked only once, which made a systematic evaluation of the entire dataset hard. The number of words per query was small. Google queries contained an average of 2.01 words in our study and HON queries 1.50 words, after removing the words image, video and media. This resulted in 191 empty queries for Google and 150 for HON. The same number of queries contained only a single character.

Removal of unclear queries

After term normalization, it became clear that there are queries unimportant for further analysis. First, a group of queries concerned sexually explicit queries: In the Google queries, the following frequent terms were removed: xxx (334 times). For HON the following terms were removed: penis (114), vagina (108), breast (102), sex (65), clitoris (32), gynecology (24). Another group of queries implicitly contained similar ideas; for Google these were: accouchement (childbirth, 143), cesarienne (33). For HON: home childbirth (239), nurse (130), birth (69). Third, another group of queries were processed to remove those not containing a precise information need, some of them, such as the term “search,” were simply placed by web robots trying to access information stored in web-accessible databases. For Google this included the following terms: medical images (508 times), HON (116), health (62), medical illustrations (32), repository (30). For HON, these terms included: search (1493), medical images (79), doctor (70), anatomy (65).

Most frequent queries and terms

After normalization and removal of queries, we analyzed the most frequent remaining terms. Table 1 shows the most frequent remaining terms forwarded from Google. This list contains very specific medical search requests, from specialists rather than patients. Most of the terms are in French, actually all of the most frequent 20. The specialized nature of the terms and the fact that they are in French can be explained with the fact that only these technical queries link towards HONmedia.

[3] http://ir.ohsu.edu/image/

[4] http://www.hon.ch/cgi-bin/HONmedia/

Table 1 - Most frequent terms forwarded from Google

Term	*Frequency*
Nerf sciatique	154
Kyste pilonidal	76
Leucemie aigue myeloblastique	72
Glossite exfoliatrice marginee	67
Fracture humerus	66
Grenouillette sublinguale	60
Hematome sous dural	57
Polype nez	56
Appendice xiphoide	53
Leucomalacie periventriculaire	51
Leucemie	46
Purpura rhumatoide	46
Scarlatine	44
Hematome retroplacentaire	40
Kyste thyreoglosse	39
Leucemie myelomonocytaire chronique	39
Leucoplasie	38
Apophyse odontoide	37
Hidradenite	37
Scoliose	34

Table 2 shows the most frequent terms directly queried with HONmedia. These terms are more likely to be from patients than specialists. The first 20 contain only a single word. More terms are in English than in French, actually all top 20, whereas a large number of the less frequent terms are in French. Most terms are of two groups: Terms describing an anatomic region or a disease. Only other terms found in the most frequent 20 are concerning symptoms or a treatment in the largest sense, such as *injection, bacteria and pain.*

Table 2 - Most frequent terms from the HONmedia search

Term	*Frequency*
Heart	381
Asthma	242
Brain	211
Diabetes	160
Liver	101
Cancer	98
Marfan	93
Kidney	77
Lung	69
Knee	69
Injection	67
Bacteria	64
Eye	60
Foot	58
Pain	58
Ear	58
Pancreas	57
Aids	57
Blood	55
HIV	54

Classified term occurrences important for us

This section analyzes only queries directly from HON as they correspond better to our needs concerning patient information search. We particularly note the most frequent terms for *anatomic region, pathology, imaging modality, symptom* and *treatment*, as these are axes to model search tasks along.

Table 3 - Frequent terms regarding modality

Term	*Frequency*
Ultrasound	47
Ecg/ekg	34/32
MRI	33
X-ray	21
Endoscopy	18

Table 3 shows modalities searched for. Interestingly, a commonly used modality (CT) is not mentioned often, whereas ECG, often discarded in medical image databases, is frequently used as it corresponds to the information needs.

Table 4 - Frequent terms regarding symptoms

Term	*Frequency*
Bacteria	64
Pain	58
Burns	42
Stress	37
Blood pressure	30

Table 4 shows symptoms searched for, where symptom is taken in a broad sense. Bacteria is not a symptom but might be interpreted from patients with flu-like symptoms looking for more information on a particular situation.

Table 5 - Frequent terms regarding treatments

Term	*Frequency*
Injection	67
Surgery	46
Stethoscope	36
Anesthesia	24
Vaccination	22

Table 5 lists terms concerning treatments, taken in a wide sense, as stethoscope is not a treatment.

Table 6 - Frequent terms regarding anatomic region

Term	*Frequency*
Heart	381
Brain	211
Liver	101
Kidney	77
Lung	69

In Table 6, frequent anatomic regions are listed that correspond well to the most frequent causes of death [14]. Also the search terms regarding pathology correspond well to diseases mentioned in [14]. Only Marfan is surprisingly frequent.

The 500 most frequent terms were analyzed accounting for almost half the search terms in total. Besides the identified five axes, some other terms are frequently queried, which are hard to classify: Human body (41), smoking (38), CPR (computerized patient record, 33), cardiology (26). It is hard to know what images or videos the users were searching for.

Table 7 - Frequent terms regarding pathology

Term	*Frequency*
Asthma	242
Diabetes	160
Cancer	98
Marfan	93
Aids/HIV	57/54

Constraints to define search tasks based on the results

From the most frequent concepts and the average number of query terms it becomes clear that users express fuzzy information needs and describe them with few terms. As the information in the HON queries corresponded better to our goal, we only used these. It is clear that information needs are often broad and it seems to aim at general illustrations (CPR, human body, AIDS …) than towards precise images of a particular modality and anatomic region. Illustrations also need to be taken into account as frequent query words such as doctor, nurse, injection or bacteria show.

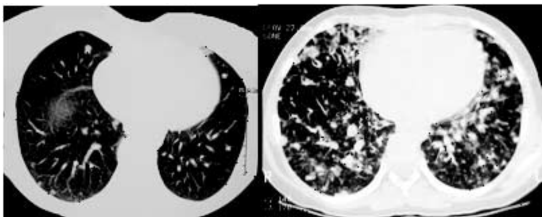
Table 8 - A collection of longer queries

Term	*Frequency*
Autonomic nervous system	16
Heart conduction system	10
Artrite reumatoide juvenil	9
Lupus vasculitis central nervous system	8
Fetal alcohol syndrome	7
Sickle cell anemia	7
Epilepsy frontal lobe	6
Respiratory distress syndrome adult	6
Spinal cord compression	6
Shoulder impingement syndrome	6

Other queries contained expected concepts but not as detailed as desired. If looking for images of the heart, all modalities, views and pathologies combined produce an extremely large number of images to be found. Such tasks are not suited to find out more about the quality of a retrieval system. For this reason, we evaluated the most frequent queries with at least three words. Table 8 lists these frequent search terms. The table shows that several terms still contain a single concept (autonomic nervous system). Most queries contain two distinct concepts, either pathology and anatomic information (epilepsy frontal lobe) or a disease and a patient group (respiratory distress syndrome adult). Still, few of these queries can be taken as query tasks for a benchmark directly.

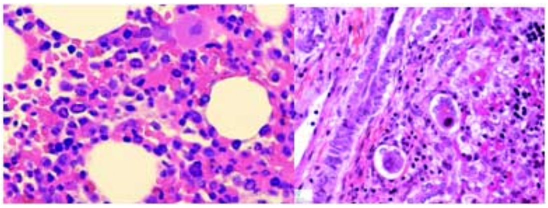
Example Query tasks of ImageCLEFmed

Finally, it was decided to use 30 real but rarer queries of the log files that cover at least two of the axes modality, anatomic region, and pathology. Example topics with example query images can be seen in Figures 1 and 2.



Show me chest CT images with nodules.
Zeige mir CT Bilder der Lunge mit Knötchen.
Montre-moi des CTs du thorax avec nodules.

Figure 1 - A visual query of ImageCLEFmed 2006



Show me microscopic images showing parvovirus infection.
Zeige mir Mikroskopien mit einer Parvovirusinfektion.
Montre-moi des images microscopiques qui montrent une infection parvovirale

Figure 2 - A semantic query for ImageCLEFmed 2006

The 30 query topics generated in this way were sent to all 40 participating research groups together with an image database. After retrieval experiments by participating groups and pooling of results, a group of physicians performed relevance judgments to compare the retrieval results of the participating retrieval systems. More about the results can be read in [15].

Discussion and conclusions

The normalization of query terms that we applied is not completely sufficient for a system that is used in several languages. A translation of the terms towards a single language or terminology would be best but with most queries being single words, this is difficult. At least 10 languages were identified. Spelling errors and abbreviations were other problems. Part of this was corrected with manual analysis but a large number of queries for the same terms could not be combined.

It can be seen that many queries for visual medical content are being performed with HONmedia search. About 52'000 queries in ten months is a large number for a small specialized search engine. Some queries are not for medical content but erotic, which is a phenomenon known by all search engines, particularly searches for images. Many queries are for illustrations of broad concepts, where the users seem to be willing to browse through a large number of varying results without a clear idea in mind and rather to illustrate an article or a presentation. Most queries are for a particular anatomic region or a certain disease. Users of the search engine do not seem to be used to supplying precise information needs concerning images. They follow the behavior of textual Internet search using broad concepts. Most image databases on the web are not well annotated and much of the information is incomplete resulting possibly in poor results.

Compared to text analysis and retrieval, medical visual information retrieval is still in its infancy. Currently, large data sets are being created and made available. Still, the applied search methods are mostly based on text, only. Techniques for visual retrieval do exist [9] and if we want to apply them in real clinical settings we need to build prototypes and make users familiar with the techniques, the possibilities and the limitations. In this sense, ImageCLEFmed is an important initiative for bringing image retrieval systems closer to routine use, through evaluating their quality. To do so, the common image databases need to be shared and realistic visual information needs have to be defined. For this, resources such as the HONmedia log files are important for us as only few medical visual search engines exist in routine use. It is also important to educate users to define their information needs more precisely using text as well as visual means and also relevance feedback.

An interesting future research topic is the analysis of query terms over short time frames. How does this behavior change with respect to events in the world (such as the bird flu)? Could the beginning of a flu outbreak be detected through keyword changes for related terms? Medical image search on the Internet and in institutional databases has a high potential but more research is needed and particularly prototypes that can be made available to the users for testing to find out more about concrete information needs.

Acknowledgements

This work was partially supported by the Swiss National Science foundation with grant 205321-109304 and by Grant 0325160 of the US National Science Foundation..

References

[1] Rice RE, Influences, usage, and outcome of Internet health information searching: Multivariate results from the Pew surveys, International Journal of Medical Informatics, 75:8-28, 2006.

[2] Gaudinat A, Boyer C, WRAPIN (Worldwide online Reliable Advice to Patients and Individuals). In MEDNET 2003, The 8th Annual World Congress on the Internet and Medicine, Geneva, Switzerland.

[3] Hersh WR, Hickam DH, How well do physicians use electronic information retrieval systems? A framework for investigation and systematic review, Journal of the American Medical Association, 1998, 280: 1347-1352

[4] Gould P, The rise and rise of medical imaging, physicsweb, 16(**8**), August 2003.

[5] Smeulders AWM, Worring M, Santini S, Gupta A, Jain R, Content-Based Image Retrieval at the End of the Early Years, IEEE Transactions on Pattern Analysis and Machine Intelligence 22(**12**) pp 1349-1380, 2000.

[6] Markkula M, Sorminen E, Searching for photos - Journalists' practices in pictorial IR. In JP Eakins, DJ Harper, J Jose (Eds.), The Challenge of Image Retrieval Newcastle upon Tyne, UK, 1998.

[7] Choi Y, Rasmussen EM, Users' relevance criteria in image retrieval in American history, Information Processing and Management 38: 695-726, 2002.

[8] Tagare HD, Jaffe C, Duncan J, Medical Image Databases: A Content-Based Retrieval Approach, Journal of the American Medical Informatics Association (JAMIA), 4(**3**):184-198, 1997.

[9] Müller H, Michoux N, Bandon D, Geissbuhler A, A review of content-based image retrieval systems in medicine – clinical benefits and future directions, International Journal of Medical Informatics, 73, pp 1-23, 2004.

[10] Clough P, Müller H, Deselaers T, Grubinger M, Lehmann T, Jensen J, Hersh W, The CLEF 2005 Cross-Language Image Retrieval Track, Springer Lecture Notes in Computer Science LNCS **4022**, pp 535-557, 2005.

[11] Hersh W, Müller H, Gorman P, Jensen J, Task Analysis for Evaluating Image Retrieval Systems in the ImageCLEF Biomedical Image Retrieval Task, Slice of Life, multimedia in medical education, Portland, OR, USA, 2005.

[12] Müller H, Despond-Gros C, Hersh W, Jensen J, Lovis C, Geissbuhler A, Medical professionals' image search and use behavior, Medical Informatics Europe, 2006.

[13] Paliouras G, Papatheodorou C, Karkaletsis V, Spyropoulos CD, Tzitziras P, "From Web Usage Statistics to Web Usage Analysis," IEEE Conference on Systems Man and Cybernetics, 1999.

[14] Anderson RN, Smith BL, Deaths: Leading causes for 2001, National Vital Statistics Report 52(**9**):1-86, 2003.

[15] Müller H, Deselaers T, Lehmann T, Clough P, Kim E, Hersh W, Overview of the ImageCLEFmed 2006 Medical Retrieval and Medical Annotation Tasks, Springer Lecture Notes in Computer Science (LNCS), 2007 – to appear.

Address for correspondence

Henning Müller (PhD)
University & Hospitals of Geneva, Medical Informatics Service
24, rue Micheli-du-Crest, 1211 Geneva 14, Switzerland
Tel +41 22 372-6175, Fax +41 22 372-8680
henning.mueller@sim.hcuge.