Journal of Digital Imaging

Analyzing Medical Image Search Behaviour: Semantics and Prediction of Query Results

Journal:	*Journal of Digital Imaging*
Manuscript ID:	JDI-14-08-0207.R1
Manuscript Type:	Hypothesis-Driven Research
Keywords:	Image Retrieval, Human-Computer Interaction, Machine Learning, Statistic Analysis, Information Storage and Retrieval
Additional Keywords:	Log File Analysis, Medical Image Search

Analyzing Medical Image Search Behaviour: Semantics and Prediction of Query Results

Abstract. Log files of information retrieval systems that record user behavior have been used to improve the outcomes of retrieval systems, understand user behavior and predict events. In this article, a log file of the ARRS GoldMiner search engine containing 222,005 consecutive queries is analyzed. Time stamps are available for each query, as well as masked IP addresses, which enables to identify queries from the same person. This article describes the ways in which physicians (or Internet searchers interested in medical images) search and proposes potential improvements by suggesting query modifications. For example, many queries contain only few terms and therefore are not specific; others contain spelling mistakes or non-medical terms that likely lead to poor or empty results. One of the goals of this report is to predict the number of results a query will have, since such a model allows search engines to automatically propose query modifications in order to avoid result lists that are empty or too large. This prediction is made based on characteristics of the query terms themselves. Prediction of empty results has an accuracy above 88%, and thus can be used to automatically modify the query to avoid empty result sets for a user. The semantic analysis and data of reformulations done by users in the past can aid the development of better search systems, particularly to improve results for novice users. Therefore, this paper gives important ideas to better understand how people search and how to use this knowledge to improve the performance of specialized medical search engines.

Keywords Image Retrieval · Human-Computer Interaction · Machine Learning · Statistic Analysis · Information Storage and Retrieval · Medical image search · Log file analysis

1. Introduction

Medical imaging studies have increased significantly in both quantity and complexity over the past 30 years [1]. Images are an essential part of medical diagnosis and treatment planning, and many tools have been created to search and interpret images, as well as to give medical doctors decision support [2,3]. Among medical specialities, radiologists are at the forefront of analyzing images, searching for specific patterns in them, and describing them in reports that form a basis for further decision making. In general, physicians increasingly use online resources to search for information. Radiologists commonly use standard search engines to look for image information for medical images [4]. Specialized radiology search engines such as ARRS GoldMiner[1], Yottalook[2] or Shambala[3] allow users to search for images in the medical literature using text queries, or in some cases, image examples to search for visual similarity. Research has shown that text search, filters for imaging modality, and image and region-of-interest search are requested by radiologists [5].

[1] http://goldminer.arrs.org
[2] http://www.yottalook.com
[3] http://shambala.khresmoi.eu

In contrast to other approaches to study users' web-site usage, search log analysis is an unobtrusive method that shows significant advantages compared to surveys and laboratory studies in scale, power, scope and location [6]. Despite limitations such as possibly imprecise user representation, less versatility, less richness, and a loose link to concepts supposed to be measured [6], search log analysis has been used in the biomedical domain to examine textual and visual retrieval systems [7].

Search logs of general search engines have been used to predict flu outbreaks and to analyze medication use [8]. They also have been used to analyze image search behaviour [9,10]. Analysis of MedLine search behaviour in the medical literature was conducted based on log files [11,12]. Closest to the presented work are the analyses of Tsikrika et al. [7] and Rubin et al. [13] that both used ARRS GoldMiner log files, but a much smaller set of queries (25,000 and 30,000 respectively, so around 10%). None of these systems performs user profiling, which would be possible with registered users of a search system. Detecting user profiles from log files was attempted in [14] but we do not try to separate queries into several user categories for ARRS GoldMiner as the technologies do not seem fully stable and our objective is to rather predict problematic queries for any user group.

Tsikrika et al. [7] analyzed 25,000 ARRS GoldMiner queries to investigate the process of query formulation and query modification in order to identify medical professionals' information needs with the aim to improve the effectiveness of the search support of such systems. This article extends the previous work using a dataset of 222,005 search queries with timestamp information. Timestamp information was not available in the previous study and was used to create user sessions with specific time limitations. Additionally, the key contribution of this paper lies in the use of machine learning algorithms to predict a query's success and the number of results for a specific query.

Similarly, Rubin et al. [13] analyzed 30,000 queries to ARRS GoldMiner and Yottalook, and implemented an algorithm for mapping search terms to RadLex[4], an ontology consisting of radiology terms, with the goal of determining what radiologists search for on the Web. As their research showed, giving the queries a RadLex semantic context improves the robustness of the analysis. Therefore, this paper also includes mapping to RadLex terms and axes, using an automatic text categorization system [15] that gives a robust mapping. This system does the mapping in three different ways, which allows to differentiate a query that is itself a RadLex term from one that includes several RadLex terms, among other cases.

The first part of the paper builds on the past work to construct a detailed analysis of a larger log file of the ARRS GoldMiner search system, while also aiming to improve technical aspects of the methodology. The second part of the paper uses machine learning to build a predictive model that is able to determine the range of the number of query results. ARRS GoldMiner retrieves all documents containing all query terms (with "AND" connection by default); additionally, if the term is in a vocabulary, the search is also done using the corresponding

[4] http://www.radlex.org

concept (MeSH, SNOMED, etc.). Therefore, it is possible to have queries with too many results and others with no results. Machine learning techniques, though widely used when working with search log files from search engines [8], have not been applied to analyze ARRS GoldMiner nor radiologists' image search behaviour [4].

The results presented in this paper provide a better understanding of the way in which physicians search for information. It also proposes two algorithms to predict whether a query will have at least one result and in what range the number of query results will be, respectively. Both algorithms have a very high accuracy and use very simple data as input, two characteristics that make them a viable alternative to be implemented in search engines as a criterion to determine when a query modification should be suggested as the computation is extremely fast. For example, if the algorithm predicts there will be too many results, the search engine could suggest the user to narrow the search; similarly, if the prediction forecasts no results, the search engine could suggest alternative queries that return results. To propose alternative queries the analysis of what other users have done in the past in terms of query reformulations, such as the one presented on this paper, can be extremely useful. For example, modifications that have been successful for other users in the past could work as a basis for suggestions made to new users. Such a recommendation system would potentially work better the more queries and query modifications it contains.

This paper is organized as follows: Section 2 includes a description of the data, of the methods used to produce descriptive analysis and of the machine learning models. Section 3 presents the descriptive analysis of radiologists' search behaviour and the results of the predictive models. Finally, in Section 4 results are discussed and Section 5 contains the conclusions.

2. Methods

2.1 Data Source

The examined query log was produced by the American Roentgen Ray Society (ARRS) GoldMiner medical image search engine [16], which currently provides access to more than 485,000 selected images from peer-reviewed biomedical journals targeted mainly to clinical professionals. The images are indexed using the keywords of the caption, the imaging modality, the age and the gender of the patient, which are all automatically extracted from the text.

The search procedure within ARRS GoldMiner always starts with a keyword search, with the possibility of filtering results at a later stage by gender, age groups, and modality. The results are returned as a set of pages, each consisting of a list of up to 10 results, or a display of up to 40 image thumbnails. Each result contains the image thumbnail, the caption, the modality and a link to the article containing the image. The acquired log file contained 222,005 consecutive queries. Each log entry included a timestamp, a client identifier (encrypted IP address to preserve privacy), the query itself and the number of results found for that query.

Preprocessing of the query logs was done in the same way as Tsikrika et al. [7]: all queries were converted to lowercase, various special characters were removed, and medical imaging modalities were normalized (for example, "XR," "X-ray" and "xray" were mapped to a single term). Consecutive identical queries in the same session and with the same number of results were considered as a single query. Such entries occur when a searcher submits a query, then views a document, and returns to the search engine. The Web server typically logs this second visit with the identical user identification and query but with a new timestamp. Also, result page navigation can cause the same logging behaviour. The log also contained identical queries in the same session that yielded different result sets; these queries were kept because they could reflect the use of filters.

2.2 Descriptive Analysis

Understanding the user's behaviour is key to enhance information retrieval systems. The first part of this paper provides descriptive analysis of the data contained in the log files.

Log analysis at session level can provide valuable information. A session is defined as a series of queries done by a single user within a small range of time where he/she attempts to fill a single information need [17]. As commonly applied, a session cut-off time of 30 minutes was defined [18]. This means that all consecutive queries within less than 30 minutes of inactivity to the previous query will be considered as a session. A query made later than the cut-off time to the previous query will be put into a new session. Query modification analysis is conducted within session boundaries and identifies the relationship between consecutive queries with three possible outcomes: query generalization, query specification and query reformulation.

In order to put the queries into a semantic context, a mapping from queries to RadLex terms was applied. RadLex is a reference ontology for the radiology domain that currently contains more than 30,000 terms used mainly for standardized indexing and retrieval of radiology information resources. It was developed by the Radiological Society of North America (RSNA) in order to satisfy needs of software developers, system vendors and radiology users by adopting the best features of existing terminology systems, while producing new terms to fill critical gaps [19,20]. Standard lexicons such as RadLex can be used to solve data-mining problems that occur due to synonyms, negation and inheritance[5]; for example, all synonyms are mapped to the same RadLex term. This mapping was mainly done to determine which of the RadLex axes were most often represented in the queries, as well as to count the term frequency of the mapped RadLex terms. The mapping from queries to RadLex terms was achieved by using Ruch's system for automatic assignment of biomedical categories [15] using lexical similarity of terms. Each term that could be mapped to RadLex was classified into one of the following 15 axes of RadLex: Imaging protocol, Report, Procedure, RadLex descriptor, Property, Anatomical entity, Imaging observation, Process, Imaging modality, Non-anatomical substance, RadLex non-anatomical set, Report component, Procedure step, Object and Clinical finding, which are the main RadLex axes.

[5] http://www.rsna.org/RadLex_in_Your_Practice.aspx

2.3 Predictive Models

A machine learning approach was applied to build a system capable of predicting the number of results a query will have. Two different tasks were defined: predicting if a query will have no results and predicting the range of the number of results (0-10 results, 10-100 results, or more than 100 results). These three classes were chosen because fewer than 10 results could be considered a query with too few results and more than 100 could be considered a very broad query where no one would look at all results, whereas in between could be considered a desirable result set.

Each query was represented by 18 attributes that were used to train the machine learning algorithms. The attributes were the following:

RadLex mappings: As explained in section 2.2, queries were mapped to RadLex terms in order to place them in a semantic context. Four types of mappings were possible: *exact* (the whole query corresponds to a term in the RadLex ontology), *all terms* (all the terms in the query can be mapped to a RadLex concept), *partial* (at least one, but not all, the terms in the query are mapped to RadLex), *none* (no term in the query can be mapped to RadLex). The first RadLex-related attribute is the type of mapping done. Given there are multiple types of mappings, each query can have between 0 and *N* RadLex mappings, *N* being the number of terms in the query. Therefore, 13 attributes were created, one for every RadLex axis present in the log files. These are binary attributes; every query is assigned a 0 or 1 in each of this variables, depending on whether the query was mapped to the axis or not.

Number of tokens in query: Two attributes were created based on the number of tokens in the query: total number of tokens and number of tokens without stopwords. The query “tumor in lung", for example, has three tokens and two non-stopword tokens.

Appearances of terms in log files: A dictionary with all the words in the queries was created, and for each of them the total number of queries in which it appears was counted. Later, this information was used to build two attributes of the vector representation of each query: *min logfile appearances* and *max logfile appearances*. In the previous example, “tumor in lung", let us assume “tumor" appears 108 times, “in" appears 2000 times and “lung" appears 520 times. Then, for this query *min appearances* = 108 and *max appearances* = 2000.

To prevent deceitful results due to unbalanced classes, the Synthetic Minority Over-Sampling Technique (SMOTE) [21] was used to balance the classes. Once this was done, random forests [22] was selected as machine learning algorithm after experiments with support vector machines [23], logistic regression [25], random forests [22] and other decision trees. The criteria used to compare them were based on correctly classified instances, kappa statistic [28], F-measure [29] and the area under the receiver operating characteristic (ROC) curve [29].

Finally, in order to analyze the impact of each attribute in the predictive model, providing understanding on which elements are relevant for prediction and which are not, an information gain attribute ranking [30] was applied to determine the importance of each attribute.

3. Results

This section describes the main outcomes of this article. In the first part, the descriptive analysis is presented. Then, the predictive models, their accuracy and other interesting metrics are exposed.

3.1 Descriptive Analysis

Terms and queries. A query corresponds to the exact text a user types into the search engine, whereas terms are extracted from the queries and might constitute the whole or part of a query. The total number of queries was reduced from 222,005 to 200,361 after preprocessing, with 92,909 queries (46%) being distinct and 75,118 queries (37.4%) appearing only a single time. In comparison to these results the study in [7], working with 25,000 records, 63% of the queries appeared a single time; the difference between these two numbers shows there is a gain in information when working with a larger dataset.

Each query was repeated on average twice, and 17,791 of the 200,361 queries (8.9%) occurred more than once. This shows that relatively few queries are repeated. The high average can be explained by the fact that the ten most frequently occurring queries represented approximately 2% of all queries. Queries that occurred only once were extremely specific terms, minor spelling mistakes that did not occur frequently, or totally off-topic queries.

Regarding the most frequently occurring terms, 33,903 (17%) of the queries contained at least one of the 10 most frequently occurring terms, and 91,589 (46%) contained one of the top 100 terms, with “cyst" being the most frequent. Figure 1 shows the proportion of queries containing the most frequently occurring terms. Tables 1 and 2 show the most frequently occurring queries and terms, respectively. Results are very similar to [7] with 7 of the most frequent queries and 9 of the most frequent terms occurring in both albeit with a slightly changing order and very different absolute numbers.

The majority of the queries consisted of two terms, followed by queries with one term, and then by those with three terms. The mean number of terms per query was 2.21; the median was 2. Among all queries, 182,004 (90.8%) consisted of three or fewer terms. In contrast, PubMed averages 3.54 terms per query [12], with a median of 3 terms per query; 80% of all queries have no more than 4 terms. Figure 2 shows the number of queries given the number of terms in it. Again, these results are very similar to results in [7].

[Here: Table 1 and Table 2]

[Here: Figure 1 and Figure 2]

Sessions. In the log files, 103,029 user sessions were identified. Among these, 100,761 (97.8%) contain less than seven queries; 64,679 (62.7%) contain only one query, 17,379 (16.9%) have two queries and 8,453 (8%) have three queries. The longest session has 126 queries.

Studying 97,315 query pairs of consecutive queries in sessions showed that, out of these, 36,056 (37.1%) do not share any common terms and only 741 (0.76%) are identical (this is influenced by result filtering), making 61,259 (62.9%) of consecutive queries in a session share at least one common term.

When analyzing the modifications done by a user in a session, 30,622 (31.4%) query pairs represent a query reformulation, followed by query generalization 16,757 (17.2%) and query specification 13,139 (13.5%). This confirms results obtained by Tsikrika et al. [7] and thus opposes the large majority of studies analyzing Web search engines logs, where reformulation is also the mostly frequently observed query modification type, but it is followed by specification and generalization [31]. Unlike Tsikrika et al. [7], available query time information allowed this study to limit the analysis to consecutive queries inside a search session, instead of all consecutive queries by the same client IP, leading to a much smaller number of query pairs relative to the search log size. According to our analysis, among the 91,375 subsequent queries in a session, the vast majority of queries 66,819 (73.1%) have a time span of less than one minute between two queries.

RadLex mapping From the 200,361 queries left after preprocessing, 124,719 (62.2%) queries could be mapped to RadLex with one of the three techniques used: 36,372 (18.2%) queries where an exact match to a RadLex concept, while 76,928 (38.4%) could be partially mapped, and 11,419 (5.7%) had every term mapped to a concept in the ontology. The remaining 75,642 (37.8%) queries could not be mapped to RadLex at all. The terms include non-medical terms spelling mistakes and terms that are too specific and not part of RadLex. In [13] 52% of the terms could be mapped to a smaller and older version of RadLex.

The most common RadLex axis is *clinical finding*, with 79,721 queries being or containing a term that could be mapped to it, which represents 40% of all queries. The second most common axis is *anatomical entity* with 38,791 (19.3%) queries, having a huge gap with the third most common axis, *RadLex descriptor*, which is only present in 22,321 (1.1%) queries (for analyzing this percentages it is very important to remember every query can be mapped to more than one or to none RadLex terms). Figure 3 shows the relationship between number of queries and Radlex axes. In [13] the most frequent axis was anatomic location (52.3%) but RadLex was much smaller at the time and it is possible that this is responsible for part of these differences with findings only covering 10.7% of the queries in this older analysis.

Among the queries, 99,060 (49.4%) are mapped to one single RadLex axis, while 23,477 (11.7%) were mapped to two axes, 2,130 (1.1%) contained terms belonging to three different axes and 52 (0.03%) to four different axes. No query was mapped to more than four axes. A similar analysis was not done in the prior work of [13].

At this point, an important question is: what axes do radiologists tend to combine for formulating their information needs? To answer this questions, the matrices in tables 3 and 4[6] show the number of times each pair of axes co-occurs. As expected, *clinical findings* and *anatomical entities*, being the most frequent axes, co-occur with others frequently. For example, the two of them co-occur in 11,787 queries, which correspond to 20% of the queries mapped to *anatomical entity*. Among the queries mapped to *RadLex descriptor*, 8,272 were also mapped to *clinical findings*, which corresponds to a 22%. The distribution of co-occurrences, however, is not only due to the frequency with which each axis appears; for *imaging observation* for example, *clinical findings* is only present in 1.9% of the queries containing it, while *anatomical entity* co-occur with it on 9.4% of its queries.

[Here: Figure 3]

[Here : Table 3]

[Here: Table 4]

3.2 Predictive Models

Machine learning algorithms were used to perform two tasks: predicting the range in which the number of results will be and predicting whether a query will or will not have results. This is a classification task, for which we aim to obtain the highest possible accuracy. Several experiments were conducted to determine which algorithm to use. In a first set of experiments, logistic regression, support vector machines (sequential minimal optimization) and random forests were tested. A model to predict the number of query results using the features based on *appearances of terms in log files* and *number of terms in query* gave an accuracy of 50.19% for logistic regression, 49.99% for support vector machines and 81.32% for random forests. This accuracy corresponds to a 10-fold cross validation using the entire dataset. Note the accuracy of random forests is lower than the accuracy finally reported, since these experiments were conducted in the first phase of the project, without taking into account the features based on Radlex mapping. Nonetheless, after finding random forests to perform radically better than the other ones, which do not even outperform the baseline (49.99% if every query is assigned to the majority class), random forests were chosen as the prefered method for the task. The default Weka[7] parameters for random forests allow the model to choose how deep each tree will be, and sets the number of trees to 10. Once the model had been trained using the whole set of features, experiments were conducted to determine if increasing the number of trees would improve the results. However, increasing the number of trees to 15 had a barely null impact on the accuracy (in the order of 10^{-3}), and therefore the final choice of algorithm uses 10 trees.

[6] CF: clinical findings, O: object, AE: anatomical entity, NS: non-anatomical substance, RD: RadLex descriptor, PP: property, P: procedure, PS: procedure step, IO: imaging observation, IM: imaging modality, RC: report component, R: report, PC: process.

[7] http://www.cs.waikato.ac.nz

The dataset used for building the model is unbalanced, which means it is not divided evenly among the classes. Therefore, after representing each query as a vector in □^{18}, the data were preprocessed with SMOTE, in order to prevent unbalanced classes in the training data from altering the results, and used to train a predictive model. To assess the performance of the algorithm, a 10-fold cross validation was used. Promising results were obtained: an accuracy of 85.19%, with an average ROC area of 0.95 and a Kappa Statistic of 0.77. More detailed information is included in table 5.

[Here: Table 5]

For predicting whether a query would have results, SMOTE was also used to balance the classes in the training data and the algorithm with the best performance was also random forests. Once again, increasing the number of trees gave almost null variation in accuracy. While the first one was a classification task between two classes, the second one classifies into three classes: 0-10 results, 10-100 results, and more than 100 results. The evaluation was also done using 10-fold cross validation and the performance is also remarkable: an accuracy of 88.29%, with a ROC area of 0.95 and a Kappa Statistic of 0.76. More details about the performance can be seen in the table 6.

[Here: Table 6]

The downside of several machine learning algorithms, such as random forests, is the low interpretability; it is hard to understand which variables are important and which are not. In order to gain insight into the role variables play in the prediction, Information Gain Attribute Ranking was used. For a class *C* and an attribute *A*, being *Ent* the entropy, the information gain, *I*, is measured by

$$I(C\ ,\ A) = Ent\ (C)\ -\ Ent\ (C\ |\ A)$$

Table 7 and 8 show the attribute's information gain for both tasks.

Given the information gain is the difference between two entropies and for each task the entropy of the class is different, the numbers cannot be directly compared (for example, the fact that in both cases *min logfile appearances* is around 35 does not mean anything). However, conclusions can be drawn from the distribution of the values, as well as for values close to zero, since these ones mean the entropy of the class and the entropy of the class given the attribute is almost the same, meaning there is no information gain from this attribute.

In both cases, *min logfile appearances* is by far the most relevant attribute. The type of RadLex mapping done to the query, the number of tokens (both with and without stopwords) and the *max logfile appearances* are important in both cases, although this last one is more relevant in the second task, which could be expected since this task also aims to predict when a query will have too many results. In both cases, RadLex axes do not provide much information.

[Here: Table 7 and Table 8]

4. Discussion

In this paper, image search behaviour of physicians and other web searchers form medical image information is analyzed based on the usage of log files, and predictive models to determine how many results a query will have are presented. The high accuracy of the predictive models, combined with the strong patterns identified in the descriptive analysis of users' behaviour, can be used to improve medical image search engines. The process of suggesting query modifications to users can be divided into two questions: when to suggest a modification and what to suggest. The findings of this paper can provide answers to both questions.

Predicting the range of the number of query results, or predicting whether a query will have results or not (depending on the desired complexity), can be used as a criterion to determine when the engine should suggest to the user a query modification. The good performance of both classifiers make them suitable candidates for being used by search engines. As these parameters are extremely simple when removing the RadLex categories, they are also extremely fast to execture, much faster than executing a query; without optimization much less than half a second could be obtained. Adding this time to a query is invisible for the user and the user can then be informed on the modifications done and the reasons for it, allowing potentially to reuse the initial query.

Once the system predicts that the query will probably not give a suitable number of results, it can make a suggestion. The information obtained from session analysis can be useful for this. Successful reformulations made by other users in the past can be used as suggestions for new users. This could be an appropriate approach whenever the query was made by another user in the past; however, as previously shown, less than 10% of the queries occur more than once, so many queries would not have a candidate for suggestion unless the log file grows massively and is available over a long period of time. Therefore, complementary methods have to be developed. The first element that can help improving a search engine is applying orthographic correction. This can reduce the number of queries with no results. As a second step, considering many searches give no results because they are too specific and others give too many results because they are too broad, it would be desirable to suggest a less or a more specific query, respectively. For the first case, a query in the log files which is contained in the current query and has obtained results could be a good candidate for a suggestion. For example, *aortitis retroperitoneal fibrosis* gives no results, so the search engine could propose the user to look for *retroperitoneal fibrosis*, which does have results. In the second case, the most common queries which contain the current query could be suggested as possible modifications. For example, if the initial input is *fibrosis*, the search engine could suggest a set of more specific queries for the user to choose from, such as *cystic fibrosis, interstitial pulmonary fibrosis, retroperitoneal fibrosis*. In this case initial results can be shown in addition to the recommended reformulations.

To further improve the results, an interesting task would be to identify off-topic queries, such as “happy new year" and “San Valentine's" that occurred in the log files. For these cases, there would be no suitable suggestion that improves the results, so the search engine could warn the user about this.

As described, the main contribution of this paper on user search log file analysis is to propose a model for medical image search engines to suggest query modifications to the users based on automatic predictions based on single queries. However, the results can also be useful for other purposes. The frequency with which certain RadLex axes appear in searches and the way in which they are combined answers the question “what are physicians looking for?". This gives valuable information to those proposing medical image retrieval tasks as benchmarks, as it is the case of CLEF eHealth [32] or ImageCLEF [33]. Knowing what radiologists or physicians in general search for is key to establishing useful tasks.

In the machine learning portion of the research, the information gain measure provides valuable insight. The fact that the most relevant attribute is *min logfile appearances* suggests there is an “offer-demand" relation, since the number of times a query has been done is useful for predicting the number of results it will get. The same happens with *max logfile appearances*.
The fact that RadLex axes are not useful for prediction is an unexpected result, since according to the hypothesis it was expected this would have impact on the number of results. However, RadLex mapping is still useful, since the type of mapping has a high information gain. This classification between the three types of mapping, part of Ruch's method [15], is particularly useful for this analysis.

5. Conclusion

This paper focuses on understanding how medical image search is performed and using this knowledge to improve specialized search engines. Data mining and machine learning techniques are applied to layout solid bases for a model of query modification suggestions. Two accurate predictive models are presented; the first one to determine when a query will have no results, and the second one to determine the range of the number of query results. In a search engine, giving no results is always a bad performance. Suggestions and modifications should be used to prevent this, and therefore predicting when it will happen is key to improving the system. The findings are promising, proving search log files can be used to train a system able to predict the level of success a search will have based on the query terms. Furthermore, a viable model that can be used by medical search engines for identifying problematic queries and modifying them to get better results is presented.

Larger log files can even improve results, since this can help to create self-learning systems. Past session information can be a valuable asset for modification suggestions to users, a field in which medical search engines still have some road ahead. In standard search engines such as Google or Bing already queries are auto-completed while typing based on past queries and their frequencies. A similar possibility exists for medical image search if sufficiently large log files are available. Even dictionaries with standard spelling mistakes can be build based on such log files. Mapping of queries to RadLex is reliable and also allows to avoid problems with synonyms as they are all mapped to a single term. Like this more can be found out on user intentions when querying, which can again be used to deliver better results than simply using key words.

Within log files, there is potentially more information that could be used to good advantage, such as click information and time spent visiting links. For Goldminer we unfortunately did not have this information available but it is again a technique frequently used in web search log files that could be transferred to medical search. The strong patterns identified in users' behaviour corroborate this is a subject that should be studied further, aiming to improve image retrieval and search engines performance for medical search. Already the described analyses potentially allows to adapt the GoldMiner system much better to the user needs by only small modifications in its functionality.

References

1. High-level Expert Group on Scientific Data. Riding the wave: How Europe can gain from the rising tide of scientific data. Submission to the European Comission, available online at http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf, 2010

2. Doi K. Computer-aided diagnosis in medical imaging: Historical review, current status and future potential. Comput Med Imaging Graph 31:198-211, 2007

3. Müller H, Michoux N, Bandon D, Geissbuhler A: A review of content-based image retrieval systems in medicine-clinical benefits and future directions. Int J Med Inform 73:1-23, 2004

4. Markonis D, Holzer M, Dungs S, Vargas A, Langs G, Kriewel S, et al.: A survey on visual information search behavior and requirements of radiologists. Methods Inf Med 51:539-548, 2012

5. Markonis D, Baroz F, Ruiz de Castaneda RL, Boyer C, Müller H: User tests for assessing a medical image retrieval system: a pilot study. Stud Health Technol Inform 192:224-8, 2013

6. Jansen BJ, Spink A, Taksai I. Handbook of research on web log analysis. IGI Global; 2009.

7. Tsikrika T, Müller H, Kahn CE Jr: Log analysis to understand medical professionals' image searching behaviour. Stud Health Technol Inform 180:1020-4, 2012

8. Yom-Tov E, White RW, Horvitz E: Seeking insights about cycling mood disorders via anonymized search logs.J Med Internet Res 16:e65, 2014

9. Müller H, Boyer C, Gaudinat A, Hersh W, Geissbuhler A: Analyzing web log files of the health on the net HONmedia search engine to define typical image search tasks for image retrieval evaluation. Stud Health Technol Inform 129(Pt 2):1319-23, 2007

10. Müller H, Kalpathy-Cramer J, Hersh W, Geissbuhler A. Using Medline queries to generate image retrieval tasks for benchmarking. Stud Health Technol Inform 136:523-8, 2008

11. Herskovic JR, Tanaka LY, Hersh W, Bernstam EV: A day in the life of PubMed: analysis of a typical day's query log. J Am Med Inform Assoc 14:212-220, 2007

12. Islamaj Dogan RI, Murray GC, Névéol A, Lu Z. Understanding PubMed user search behavior through log analysis. Database (Oxford) 2009:bap018, 2009

13. Rubin DL, Flanders A, Kim W, Siddiqui KM, Kahn CE Jr: Ontology-assisted analysis of web queries to determine the knowledge radiologists seek. J Digital Imaging 24:160-164, 2011

14. Palotti J, Hanbury A, Müller H, Exploiting Health Related Features to Infer User Expertise in the Medical Domain, Web Search Click Data workshop at WSCM, New York City, NY, USA, 2014.

15. Ruch P. Automatic assignment of biomedical categories: toward a generic approach. Bioinformatics 22:658-664, 2006

16. Kahn CE Jr, Thao C: GoldMiner: a radiology image search engine. AJR Am J Roentgenol 188:1475-1478, 2008

17. Silverstein C, Marais H, Henzinger M, Moricz M: Analysis of a very large web search engine query log. SIGIR Forum 33(1):6-12, 1999

18. Jones R, Klinkner KL. Beyond the session timeout: automatic hierarchical segmentation of search topics in query logs. In: Proceedings of the 17th ACM conference on Information and knowledge management. ACM; 2008. p. 699-708.

19. Langlotz CP: RadLex: a new method for indexing online educational materials. RadioGraphics 26:1595-1597, 2006

20. Rubin DL: Creating and curating a terminology for radiology: ontology modeling and analysis. J Digit Imaging 21:355-362, 2008

21. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research. 2002;16:321-357.

22. Breiman L. Random forests. Machine Learning 45:5-32, 2001

23. Chang CC, Lin CJ. LIBSVM: a library for support vector machines; 2001.

24.Kohavi R. The power of decision tables. In: Machine Learning: European Conference on Machine Learning-95. Springer; 1995. p. 174-189.

25. Le Cessie S, Van Houwelingen J. Ridge estimators in logistic regression. Applied Statistics. 1992;p. 191-201.

26. Holmes G, Pfahringer B, Kirkby R, Frank E, Hall M. Multiclass alternating decision trees.
In: Machine Learning: European Conference of Machine Learning 2002. Springer; 2002. p. 161-172.

27. Quinlan RJ. C4.5: Programs for Machine Learning. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.; 1993.

28. Viera AJ, Garrett JM: Understanding interobserver agreement: the kappa statistic. Fam Med 37:360-363, 2005

29. Manning CD, Raghavan P, Schütze H. Introduction to Information Retrieval. Cambridge University Press, 2008

30. Hall MA, Holmes G. Benchmarking attribute selection techniques for discrete class data mining. IEEE Transactions on Knowledge and Data Engineering 15:1437-1447, 2003

31. Hollink V, Tsikrika T, de Vries AP: Semantic search log analysis: a method and a study on professional image search. J Am Soc Inform Sci Tech 62:691-713, 2011

32. Goeuriot L, Kelly L, Li W, Palotti J, Pecina P, Zuccon G, et al. ShARe/CLEF eHealth Evaluation Lab 2014, Task 3: User-centred health information retrieval CLEF eHealth overview. In: CLEF Proceedings. Springer LNCS; 2014.

33. Seco de Herrera AG, Kalpathy-Cramer J, Demner Fushman D, Antani S, Müller H, Overview of the ImageCLEF 2013 medical tasks, CLEF working notes 2013, Valencia, Spain, 2013.

[1] http://goldminer.arrs.org/

[2] http://www.yottalook.com/

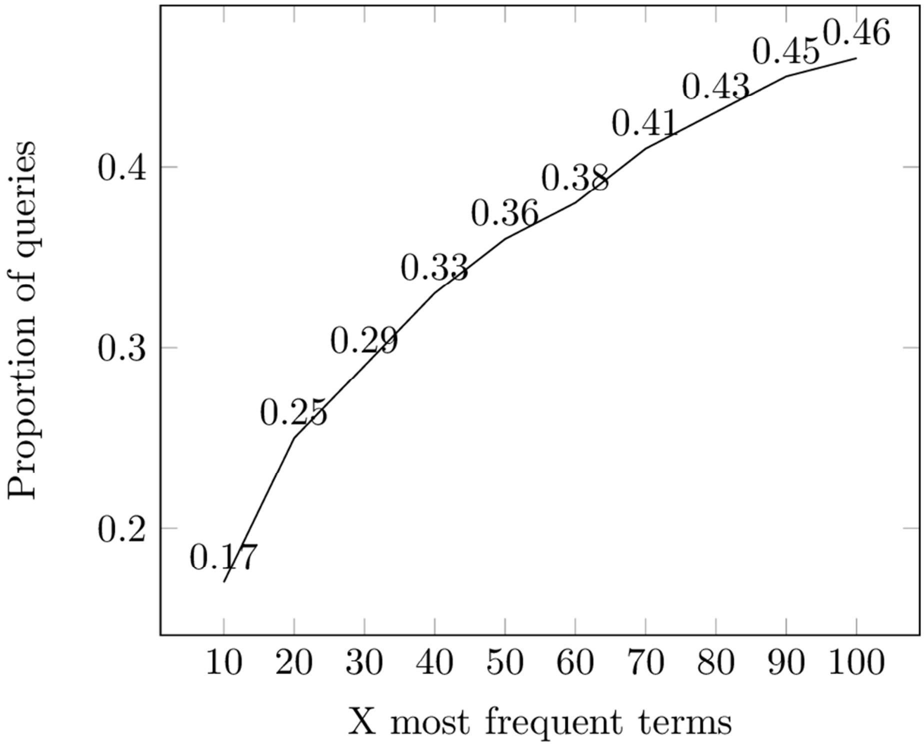
[3] http://shambala.khresmoi.eu/

[4] http://www.radlex.org/

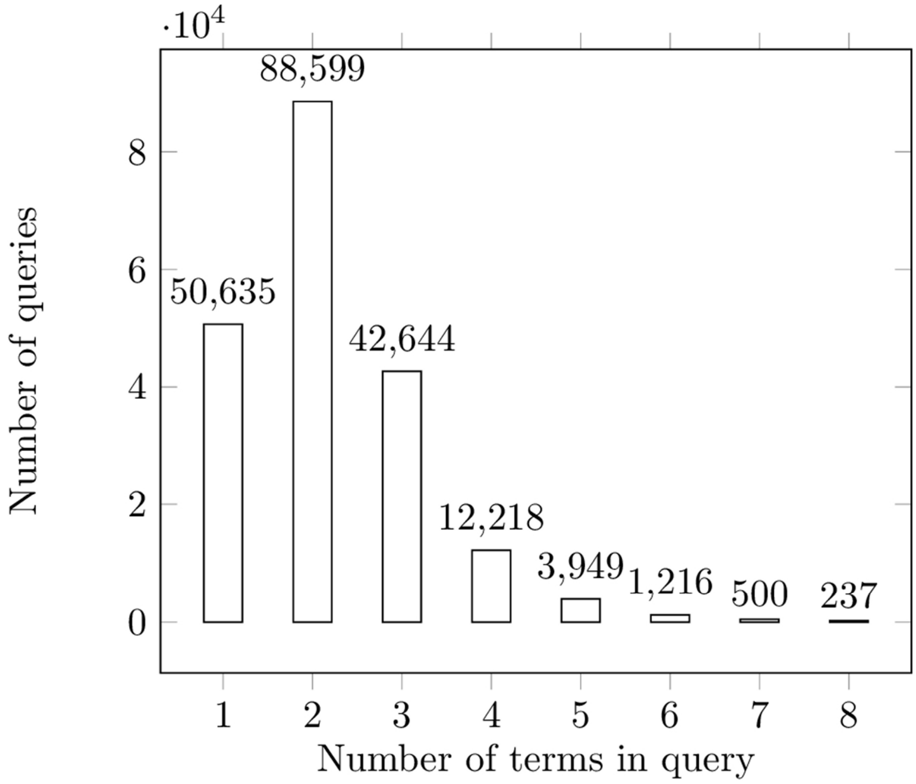
[5] http://www.rsna.org/RadLex_in_Your_Practice.aspx/

[6] CF: clinical findings, O: object, AE: anatomical entity, NS: non-anatomical substance, RD: RadLex descriptor, PP: property, P: procedure, PS: procedure step, IO: imaging observation, IM: imaging modality, RC: report component, R: report, PC: process.

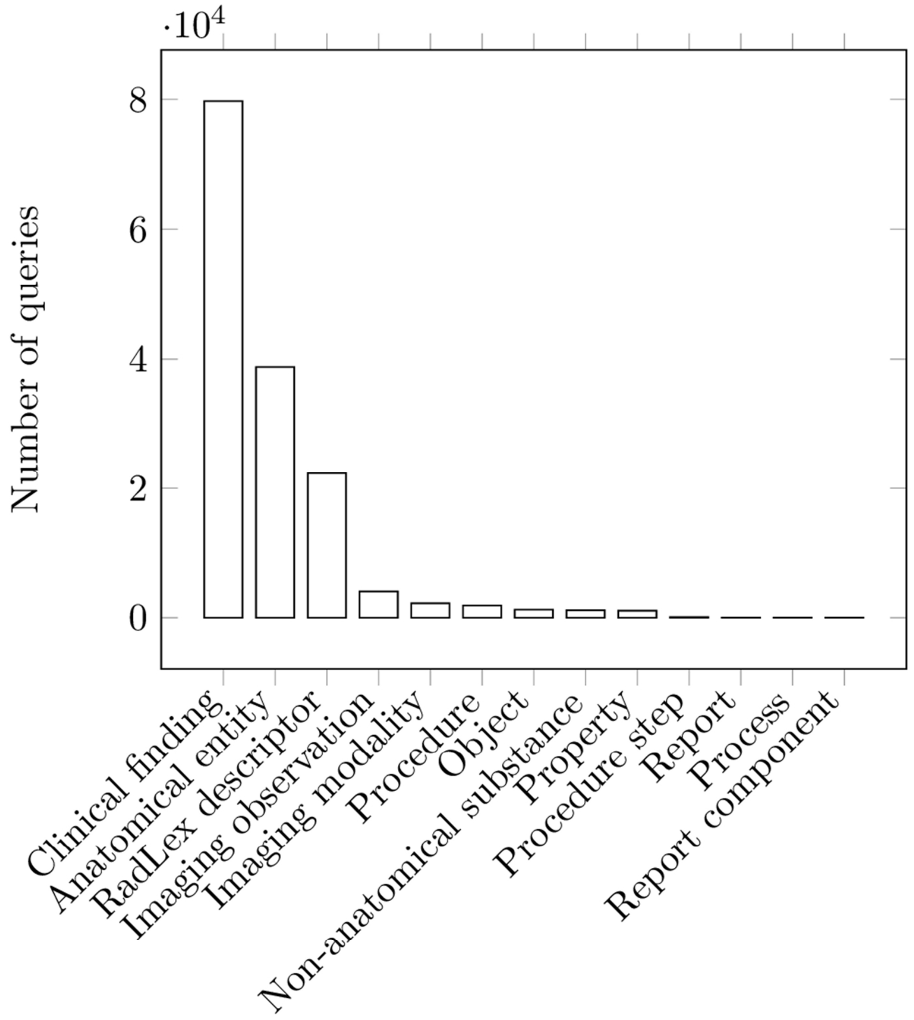
[7] http://www.cs.waikato.ac.nz/



Proportion of the queries containing the most frequently occurring terms.



The number of queries with a specific number of terms in the query.



Number of queries mapped to each RadLex axis.

	Query	Frequency
1	mega cisterna magna	820
2	baastrup disease	798
3	limbus vertebra	462
4	toxic	428
5	cystitis cystica	405
6	buford complex	274
7	thornwaldt cyst	274
8	splenic hemangioma	254
9	double duct sign	249
10	cystitis glandularis	245

	Term	Frequency
1	cyst	6346
2	mri	3536
3	disease	3536
4	ct	3504
5	fracture	3366
6	tumor	3233
7	syndrome	2994
8	liver	2486
9	pulmonary	2424
10	sign	2293

	CF	O	AE	NS	RD	PP
CF	79721	175	11787	150	8272	225
O	175	1243	229	4	89	7
AE	11787	229	38791	116	5217	166
NS	150	4	116	1161	55	7
RD	8272	89	5217	55	22321	18
PP	225	7	166	7	189	109
P	280	18	357	4	163	16
PS	0	1	12	0	1	0
IO	97	6	488	2	543	16
IM	552	25	580	2	249	9
RC	2	0	5	0	3	0
R	4	0	1	0	0	0
PC	1	1	5	0	0	0

	P	PS	IO	IM	RC	R	PC
CF	280	0	97	552	2	4	1
O	18	1	6	25	0	0	1
AE	357	12	488	580	5	1	5
NS	4	0	2	2	0	0	0
RD	163	1	543	249	3	0	0
PP	16	0	16	9	0	0	0
P	1889	1	11	23	0	1	0
PS	1	101	0	0	0	0	0
IO	11	0	4044	12	0	0	0
IM	23	0	12	2211	0	0	0
RC	0	0	0	0	10	0	0
R	1	0	0	0	0	16	0
PC	0	0	0	0	0	0	12

	R1	R2	R3	Weighted Av.
Precission	0.842	0.819	0.874	0.85
Recall	0.876	0.688	0.92	0.851
F-measure	0.899	0.748	0.897	0.849
ROC Area	0.955	0.92	0.971	0.953

	# Res>0	# Res<0	Weighted Av.
Precision	0.899	0.865	0.884
Recall	0.899	0.864	0.884
F-measure	0.899	0.865	0.884
ROC Area	0.951	0.951	0.951

Variable	Info.Gain
min logfile appearances	0.35278316
Type of RadLex mapping	0.10706495
max logfile appearances	0.09828245
number of tokens	0.07604782
number of non-stopword tokens	0.07554565
RadLex: clinical finding	0.02718913
RadLex: non-anatomical substance	0.00130726
RadLex: imaging observation	0.00129999
RadLex: anatomical entity	0.00082734
RadLex: procedure	0.00047458
RadLex: property	0.00042359
RadLex: RadLex descriptor	0.00035407
RadLex: imaging modality	0.00033401
RadLex: object	0.00026038
RadLex: procedure step	0.00016858
RadLex: process	0.00001056
RadLex: report component	0.00000342
RadLex: report	0.00000335

Variable	Info. Gain
minlogfileappearances	0.3625514
maxlogfileappearances	0.1735592
numberofnon-stopwordtokens	0.1498272
numberoftokens	0.1497191
TypeofRadLexmapping	0.1130494
RadLex:clinicalfinding	0.0122519
RadLex:RadLexdescriptor	0.0091736
RadLex:imagingobservation	0.0018093
RadLex:property	0.0016000
RadLex:non-anatomicalsubstance	0.0013986
RadLex:anatomicalentity	0.0013594
RadLex:imagingmodality	0.0009119
RadLex:object	0.0006390
RadLex:procedure	0.0001619
RadLex:procedurestep	0.0001126
RadLex:report	0.0000384
RadLex:process	0.0000363
RadLex:reportcomponent	0.0000165

Table/Figure	Legend
Table 1	The most frequent queries in the logfile.
Table 2	The most common terms occurring in the queries.
Table 3	Co-occurrence of RadLex axes in the queries (first part containing CF, O, AE, NS, RD, PP).
Table 4	Co-occurrence of RadLex axes in the queries (second part containing P, PS, IO, IM, RC, R, PC).
Table 5	Performance of Random Forests for predicting if a query will have results or not.
Table 6	Results of Random Forests for predicting the range of the number of query results. R1 has less than ten results (including no results), R2 has between 10 and 100 results, and R3 has more than 100 results.
Table 7	Relative influence of variables for predicting if a query will have no results, according to Info Gain Evaluation.
Table 8	Relative influence of variables for predicting the range of the number of query results, according to Info Gain Evaluation.
Figure 1	Proportion of the queries containing the most frequently occurring terms.
Figure 2	The number of queries with a specific number of terms in the query.
Figure 3	Number of queries mapped to each RadLex axis.