Grant Agreement Number: 257528

KHRESMOI

www.khresmoi.eu

Report on automatic document categorization, trustability and readability

Deliverable number	*D1.6*
Dissemination level	*Public*
Delivery data	*due 31.8.2013*
Status	*Final*
Authors	*Allan Hanbury, Célia Boyer, Ljiljana Dolamic, João Palotti*

This project is supported by the European Commission under the Information and Communication Technologies (ICT) Theme of the 7th Framework Programme for Research and Technological Development.

Executive Summary

This deliverable reports on the update of the automatic categorization and present advances on the readability and trustability detection in the task T.1.5 *Categorization, Trustability and Readability* in the KHESMOI WP1 *Large Scale Biomedical Text Mining and Search*. In this report we describe and evaluate the tools developed in the scope of this work package. The document categorization tool presented here and its implementation within the search engine enables the end user to filter the results returned by the search engine according to the set of predetermined set of classes, such as Cancer or Alcohol.

Taking into account the fact that not all users have the same level of medical literacy, two differant approaches have been taken in order to propose to the end user the suitable documents. Firstly, classification of documents by the readability level has been performed. Documents have been labeled as being either easy or difficult to read. Integration of this tool within the search engine gives to the end user possibility to access only documents he judges suitable. Secondly, a method for prediction of user's expertise has been proposed. The goal of this method being the ability to present the user with the documents addapted to his/hers literacy level.

Can the user have confidence in the information from the pages returened by results of the search engine is another issue we have tackled in this research. With the goal of determining the level of trustability, a tool based on machine learning, whose results are presented here, have been developed and integrated in the crawling process. Integration of its results into the search engine will enable the visualisation of te level of trust aassigned to the source by the system.

Table of Contents

List of Figures

List of Tables

Abbreviations

DF	document Frequency
HON	Health On The Net
HONCode	HON Code of Conduct
MeSH	Medical Subject Headings
NB	Naive Bayes
NLM	National Library of Medicine
NLTK	Natural Language Tool Kit
SVM	Support Vector Machine
TRIP	Turning Research Into Practice
ZS	Z-score

1 Introduction

Medical information, especially if it is made available to general public, needs to be reliable and trustworthy. Taking into account the overwhelming quantity of medical information currently available, it has become very difficult for users, especially for patients, to find medical information they can judge trustworthy [18], [6]. Health On The Net (HON) has established a third party accreditation program, HON Code of Conduct (HONCode), with a goal of helping patients in this direction. In order to identify if a candidate website clearly indicates the information required by HONCode principles, the experts at HON perform a manual inspection. Taking into account the ever increasing number of online documents, this approach, even though very reliable, becomes highly inefficient. This creates the need for an automatic way of performing such a task.

Moreover, in order to be correctly appreciated by the end user, the information needs to be adapted to the users' capability. On the Internet, easy-to-read health documents coexist with technical and scientific health documents. A difficult-to-read text could lead people to give up reading it or even to misunderstand the content. As shown by various studies in diverse areas of medicine [10, 9], the high level of readability required by health websites may make comprehension difficult for a substantial portion of the general public. In our research we aim to help the user to find the most appropriate documents according to his/her level of literacy.

Therefore, it is also necessary to automatically classify users according to their level of expertise. This can assure an even richer experience for end users, as we could re-rank the search results to cope with different levels of readability, for example.

The remainder of this deliverable is structured as follows. Section 2 focuses on the document classification, assigning for each document a topic (Section 2.1), a trustability score (Section 2.2) and a level of readability (Section 2.3). Section 3 shows how to make use of an extended log analysis (Section 3.1) to develop a classifier for user expertise (Section 3.2). We conclude and present the future work in Section 4.

2 Document Classification

2.1 Topics

Basic description on the categorization work performed at HON and its integration within the Khreshmoi Classic Search System are given in the deliverable 8.3 [13], Section 3.1.3. In this section we present the detailed algorithm, description of the data used for the evaluation of the developed system as well as the results of the performed experiments. .

2.1.1 Data

Key phrases acquisition In order to use the described algorithm on the set of pre-requested classes we needed to create a list of key-phrases related to each one of them. Since the classes belong to the public health domain, both medical and general topics are covered.

Using existing medical dictionaries such as MeSH has proven to be a bad solution for two reasons: (1) while coverage of certain topics such as "cancer" is very good, public webpages regarding cancer tend to use a vocabulary which does not correspond to the one used in MeSH. (2)

the coverage of other "lesser medical" topics is fairly poor. Another possible solution would be manually creating lists using domain experts, however it is both expensive and time consuming.

We decided to automatically generate the lists of key phrases per topic, based on the content of publicly-available glossaries from trusted public health sources (HONcode-certified web sites). The full list of the glossarie entries such as "Glossary_of_Smoking_Cessation_Terms" was used to create the initial list of phrases. The list of phrases created in this way were then manually examined in order to remove the candidate terms that were too general and thus not representative of the given public health topic. For example, we can take the key phrase "Orphan drugs" as being a part of the glossary for the topic "Rare diseases", while none of the separate components of this phrase would be considered as a good indicator of this topic. The number of the retained key-phrases varies from one class to another and spans from 16 for "Statistics" to 191 for "Rare Diseases" class. The length of the key-phrases spans from 1 to 5 words.

Algorithm For each class a list of related key-phrases is created as described in previous paragraph. The document d is considered to be an ordered list of words $\{w_1, w_2 \cdots w_{|d|}\}$ where w_j is the j^{th} word in the document while $|d|$ is the length of the document. If we assume that the class C has a key-phrase list $KP_c = \{kp_1, kp_2 \cdots kp_{|k|}$ where $|k|$ is the length of the list, the algoritham works as follows:

- For each key-phrase kp_j in the KP_c the number of its appereances is determined within the document in the following manner:
  - All permutations of the linguistically treated key-phrases (stemming and stopword removal) are matched on a larger sliding window within the document d. The size of the window depends on the key-phrase length $|kp_j|$ and spans form $|kp_j|$ to $4*(|kp_j|-1)+1$. This means that for a two word long phrase the max tested window would be five words long. Thus the key-phrase $kp_j = \langle kt_1 \rangle \langle kt_2 \rangle$ would match either $[\cdots]\langle kt_1 \rangle [w_{i-1}][w_i][w_{i+1}]\langle kt_2 \rangle [\cdots]$ or $[\cdots]\langle kt_2 \rangle [w_{i-1}][w_i][w_{i+1}]\langle kt_1 \rangle [\cdots]$ in the document d.
- The weight of the key-phrase kp_j within the document d is calculated as follows: $w_{kp_j} = |kp_j| * n_{kp_j}$ Where n_{kp_j} represents the number of appearances of the phrase kp_j in the document d.
- The density of the key-phreses fro the set KP_c in the document d is: $d_c = \sum_{j=1} |k| \frac{w_{kp_j}}{|d|}$
- The document d is classified into the class C if the value of d_c is higher than an empirically determined threshold.

Collection The collection used for the evaluation of the presented algorithm was created from the public health related web pages. Double blind manual classification into the set of prerequested classes was used.The list of clasess has been established by HON based on user requirenments. They have been listed in the Section 3.1.3 of the deliverable 8.3 [13]. Only pages having the concordance of both judges were kept. A total of 530 pages were retained in this process to be used in the algorithm evaluation.

Taking into account the characteristics of the documents, the number of classes the page could be placed in was not limited. Thus we might have in the collection the pages being classified to more than one class. For example, a page about influence of tobacco and alcohol to cancer development would be classified into three classes: "Tobacco", "Alcohol" and "Cancer".

2.1.2 Results

Table A.1 shows the values of precision (P), recall (R) and F-measure (F) respectively for selected set of classes regarding different values of the density threshold. In these tables, the precision provides the ratio of correctly assigned positive class among all positive predictions (precision). The recall provides the ratio of positive examples correctly predicted (recall). While the F-measure, also known as F_β measure [28] is a combined effectiveness measurement. Calculated according to equation 1 this measure gives an equal importance to the precision (P) and recall(R) when β is set to 1, which is the case in our results.

$$F_\beta = \frac{(\beta^2+1)PR}{\beta^2 PR} \tag{1}$$

The set of classes for which the results are presented in this section is selected in order to ilustrate the classification performance. As it can be noticed, the value of the precision increases as the density threshold is set to a higher value. For example in the case of the "Cancer" class while the precision value is 0.5349 when the threshold is set to 1 it reaches 0.9525 for a threshold of 6 and 1 when the threshold is set to 10. In our system the threshold is set to 3, which proves to be the best solution in terms of precision-recall trade-off.

2.1.3 Search engine integration

Integration of this tool into the content extraction process and the search engine is described in deliverable 8.3 [13], Sections 3.1.3 and 3.2.

2.2 Trustability

2.2.1 Data

When it comes to automatic text categorization the most important component of the system is the knowledge base containing positive and/or negative examples. When performing the evaluation whether the website is respecting or not certain HONCode criteria, human experts were asked to extract the text demonstrating the respect of the given criterion. This way we have obtained the knowledge base related to each of the 8 HONCode principles: *Authority*, *Complementarity*, *Privacy*, *Attribution*, *Justifiability*, *Authorship*, *Sponsorship* and *Advertising*.[1] Here are the examples of the extracts taken by the experts for two criterias (transformed into lower case):

- Complementarity:
  "the information that we provide on our web site is designed to support, not replace, the

[1] More information and details on the HONCode principles can be found at http://www.hon.ch/HONcode/Conduct.html

relationship that exists between a patient/site visitor and his/her physician. please keep in mind that the text provided is for informational purposes only and is not a substitute for professional medical advice, examination, diagnosis or treatment. always seek the advice of your physician or other qualified health professional before starting any new treatment or making any changes to existing treatment."

- Privacy:
  "privacy policy this web site does not collect information from any visitor. cookies are not used at any trime. we do not collect email addresses and any communication will not result in the retention of your information in any form. we do not keep a database of visitor information or any other statistics regarding the demographics or other attributes of users of this web site. there are opportunities to become a patient of the norman endocrine surgery clinic, however, this occurs on two specific pages designed for this purpose. these two pages are hosted on a secure server. you will know that you are entering your data and warnings will be given. this is a clear decision that you will make. these two pages are encrypted and secure and are clearly marked as such. these two pages have the logos and secure certificates clearly visible. the information entered on these two secure pages is not accessible to anybody except the medical staff of the norman endocrine surgery clinic. these two secure pages have been approved and meet all current 2004 standards for secure online medical information as established by the american medical association."

The principle *Attribution* has been divided into two separate criteria namely *Reference* and *Date* due to different requirements for these two elements within a single criterion. Table A.2 shows the number of extracts available at present for each of the criterion in various languages. However the HON study and its implementation into this project is limited only to English. The training data for this language can be considered to be the most complete one, as it can be seen from the table.

Each extract obtained in this manner represnets one document within the training/test collection.

2.2.2 Methods

Unlike the previous studies conducted by HON in the domain of automatic detection of the HONCode principles [12], in this research we are using the whole document as the classification unit, believing that the document is more suitable for this purpose, since the statement about a certain criterion is spread within the whole document, and not concentrated into a single sentence. Since a document can conform to more than one criterion, we are facing the classification of the text into the classes that are not mutually exclusive (*any-of classification*). It means that if a document conforms to one criterion, it does not imply that it conform (or disconform) to any other. We are taking the following approach described in [19]:

1. Build a classifier for each class, where the training set consists of the set of documents in the class (positive labels) and its complement (negative labels).

2. Given the test document, apply each classifier separately. The decision of one classifier has no influence on the decisions of the other classifiers.

We ran all the experiments using 10-fold cross-validation. Also, the documents were treated linguistically: the stopwords were removed and the Porter stemmer was applied.

Machine learning algorithms The machine learning algorithms used in our system are proposed by the learning framework [32]: Naive Bayes (NB) and Support Vector Machines (SVM). We have applied these two algorithms using various features types and feature selection algorithms on our English collection.

Features In all our experiment we have treated the documents linguistically, removing the stopwords and applying Porter stemming. The learning unit was then set to one of the following:

1. single word (W1) - "privacy information" → "privat inform"
2. word co-occurrence (COOC) - "privacy information" → "privat_inform"
3. character n-gram, in our case 4-gram (C4) - "privacy information" → "priv riva ivat info nfor form "

The single word W1 (bag of words) has been chosen as a baseline of this research. The choice of the n-gram was driven by the fact that this type of tokenization [20] has proven to be effective for different languages, especially in the case of lack of the linguistic tools such as stemmers [21] they have also proven to be more robust to typos[25].

Features selectors In order to reduce the feature space dimension and to distinguish the features that could help us determine the document's class, we have opted following two criteria:

1. Document Frequency (DF) - favors features distributed in largest number of documents. In our research this type of feature selection is used as a baseline.
2. Z-score [33] - favors the features overused within the training collection. This method supposes that the terms belonging to general vocabulary would have a Z-score between -2 and 2, a word with Z-score greater than 2 would be concidered as being overused within the class.

We have then used various levels of features reduction, keeping 30%, 50% or 80% of features with a goal of demonstrating how the reduction of the number of features influence the classifier performance.

2.2.3 Results

Tables A.3, A.4 and A.5 give the obtained experimental results for the nine criteria previously listed.

The results displayed in this section, present a part of all results obtained during this research. They have been chosen based either on the performance obtained for the particular combination of parameters, or for illustrating the influence of the certain parameters on the obtained result. The best result for the three measurements are shown in bold.

As it can be seen from the presented results, the choice of the best performing parameter combination depends largely on the criterion. For example, while the best recall for "Privacy" (Table A.3) is achieved by the combination of {NB, COOC, DS, 0.8}, the same combination was far from the optimal for recall, precision and F-measure for "Authority". The choice of the correct parameters is also driven by the need of maximization of precision, recall or F-measure.

Based on the results obtained during this research, the parameters mentioned above are chosen for each of the criterion to be incorporated within the page crawling process as well as the search engine integration.

2.2.4 Search Engine Integration

Each page crawled by the HON crawler described in the [13] is tested for each of nine criteria presented here. This information is kept in both CouchDB (described in details in [13]) , as well as indexed into the search engine. Based on the information gathered in this process, on all pages crawled from a domain, the level of trust of the website is determine and will be displayed. Even though this feature is not yet integrated within the search engine interface, it has been planned to be displayed with the search result in the form of an icon indicating the level of the trust detected for the page domain.

2.3 Readability

2.3.1 Description

For the goal of determining how difficult to read are the medical documents, HON has taken two main research and development approaches. In this stage we have defined two categories for the document readability level namely easy and difficult, althought more intermediate categories could be added in the future. As a first approach HON has used the machine learning algorithms described in previous section. Different combinations of features types, features selectors and categorization algorithms were applied to both English and French data.

The second approach used is the lexical approach, tested only on the English data due to the lack of corresponding lexical data for the French.

2.3.2 Test collection

For the English language, the collection used in this evaluation contains 592 pages from four sources accredited within the HON search engine [11], namely:

1. *HONnews* - collection of medical news from HON and HealthDay web sites (101 page)
2. *MedlinePlus* - medical pages for non-expert users (101 page)
3. *OESO* - scientific articles about Oesophagus diseases (289 pages)
4. *eMed* - documents for experts from *eMedicene* site (101 page)

In this collection the articles sourced from *HONnews* and *MedlinePlus* are considered to be easy to read, while the remaining two sources, addressing the field experts are concidered to be difficult to read for a non expert.

The French language test and training collection consists of following documents:

1. *CISMEF* - Catalogue et Index des Sites Médicaux de langue Française [2](541 page)
2. www.passeportsante.net
   (a) Plantes (218 pages)
   (b) Therapies (124 pages)
3. Vidal (231 page)

The pages coming from CISMEF are the ones in this collection considered to be difficult, while the rest of the sources provide easily readable content.

These collections were used for both lexical and learning methods of readability detection and evaluation. For the lexical part of this development, performed only for English, the lexical data, describe in the following section was also used.

Lexical data Previously mentioned lexical material required for the evaluation of the technical difficulty of the given medical text was extracted from freely available lexical sources. Two types of vocabularies were used in this study. Common vocabularies composed of lemmas and their forms:

1. *YAWL* list, over 264 000 entries
2. *ENABLE* word list 173 000 entries
3. *5desk* list compiled from 5 monolingual desktop dictionaries, over 120 000 entries
4. *12dicts* lexicon compiled from 12 bilingual dictionaries, nearly 82 000 entries

As the source of medical terminology the MeSH vocabulary was used, providing 22 900 simple (single term) and complex (multiple terms) main descriptors.

2.3.3 Methods

As mentioned before we have used two methods for determining the documents' readability level. This section gives a short description of the methods used, while the obtained results are given in the next section.

Learning In the learning part of our experiments we have combined various features types, features selection algorithms and feature reduction levels with NB and SVM learning algorithms (described in 2.2). The goal of the performed experiments was to find the best performing combination of these parameters, to be implemented into the HON search engine. We ran 5-fold cross-validation, using 80% of French documents for training and the remaining 20% for testing. The results of these experiments are given in Table A.7. This method has also been tested on our English collection but the obtained results have led us to the conclusion that the collection for this language is to homogeneous to be used for this purpose (e.g. precision 1 for cross validation). The additional test will be performed on a newly aquired collection for English with a goal of clarify the problems encountered with the current collection.

[2]www.chu-rouen.fr/cismef/

Lexical For English we have opted for the lexical method of readability detection, since it has proven to be effective for this language. This method was originaly presented in [2], with a slight correction in the used formula. The following algorithm based the lexical data described in 2.3.2 is used in the experiments presented in this section:

- The following stop word list were removed from the document text: the, of , and, to, a, in, that, is, was, he, for, it, with, as, his, on, be, at, by, I , a, b, c, d, e, f $\cdots$ z
- Terms within the lexical files are weighted as follows:
  - If term belongs only to the MeSH dictionary $\rightarrow w_t = 100$
  - If the term appears in easiest 12dicts category $\rightarrow w_t = 1$
  - Others $\rightarrow$ these terms are wighted with the assumption that the less complex terms appear more frequently within the corpora, in this case the lexical data.
- weighting
  - word score: $w_s = w_t * n_t$ - where w_t is the term weight previously described while n_t is given by:
    $n_t = m * (\frac{a}{f^b})$ - where:
    - For the runs whose results are presented in this report the multiplier was set to $m = \frac{w_t}{10}$.
    - $a = 1, b = 1.5$
    - f - term frequency (number of term appearances in the document)
  - sentence score: $s_s = \frac{\sum w_s}{\sum n_t}$
  - document score: $d_s = avg(s_s)$
- The threshold value is determined empirically. Documents having the score d_s below the threshold are seen as being easy to read while those above this value are designated as difficult.

2.3.4 Results

Table A.6 shows the values of precision, recall and F-measure for different thresholds. As it can be seen from the table, the threshold should be set around 33 in order to achieve the best precision results.

2.3.5 Search engine integration

The readability level detection has been integrated in the content extraction process of the HON search engine, described in details in the [13]. It has also been integrated within the search interface, giving the information on the readability complexity of the returned document by displaying a flag next to the search result as illustrated in figures 1 and 2.

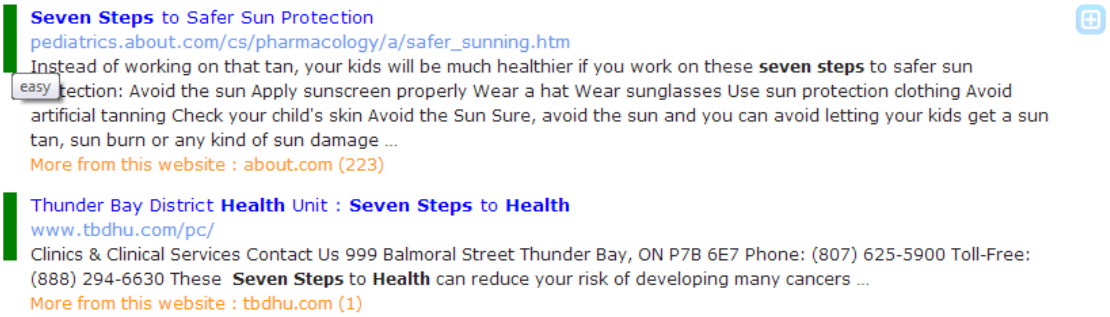


Figure 1: HON Search result labeled as easy to read

Figure 2: HON Search result labeled as difficult to read

3 User Categorization

In this section, we present a method to predict the user expertise based on his/her issued queries. This would allow the system to better support interations with users, providing documents at the appropriate level of the user expertise.

Among the methods available to study user searching behaviour, one of the most effective is the analysis of user interactions logged by search engines [16]. The use of logs is unobtrusive and captures the user behaviour in a natural setting. Therefore, we employ transaction logs from diverse sources to analyse how the users search for medical content.

In Section 3.1 we show the data used, the processing steps and the analysis made. Section 3.2 presents the classifier built to classify users according to their expertise and Section 3.3 discusses the results and future directions.

3.1 Log Analysis

This section is dedicated to the logs analysis. We describe the datasets and all the preprocessing steps applied to them in Section 3.1.1. In Section 3.1.2, we analyse in depth the search logs used, from general statistics, characterizing the logs, to the semantic focus of a session, representing the user intent when searching for health information on the Web. All these analyses are used to generate the features for the classifier described in Section 3.2.

3.1.1 Data and Preprocessing Steps

Data We utilize four query logs divided into five datasets in our analysis: two focused on laypeople queries, two made of queries from professionals and one consisted of queries not related to health or medical information.

The query logs assumed to consist almost completely of queries submitted by laypeople were obtained from health-related searches in America Online's search service [26][3] and from the Health on the Net Foundation website (HON[4]).

The AOL logs were obtained from March to May of 2006 and divided into two non-overlapping sets: **AOL-Health** and **AOL-NotHealth**. For this purpose, the click-through information available in the AOL data was used. Whenever a clicked website was found among the pages classified in the Health category of the Open Directory Project (ODP)[5], the query was considered health related. Unfortunately, there is a considerable number of queries without click-through data (47%). For these queries, we considered them as health-related if and only if there is an identical query previously classified using the ODP data. All other queries were judged as not related to the health or medical domain, forming the AOL-NotHealth dataset. The final processing step in the AOL data consisted of the manual removal of some misleading queries from AOL-Health , such as "jobs" or "yahoo". This last query, for instance, was present in the AOL-Health logs because *http://health.yahoo.com* is in the Health category of ODP, but only "yahoo" is a very generic query and not necessarily related to health. Another important query removed is the single dash ("-"). It is the most popular query in the whole AOL log, covering a wide range of unrelated visited results (from pubmed to pornographic sites), most likely used to mask the identity of users [5].

The **HON** dataset is composed of anonymous logs ranging from December 2011 to February 2013 and collected through the search engine used to access the HONCode certified sites [3]. Although the majority of the queries are issued in English, the use of French or Spanish is very frequent. Aiming to reduce noise, only queries consistent with Unicode block Latin 1 (iso-8859-1) were kept[6].

As professional datasets, we are using the logs from the Turning Research Into Practice (**TRIP**) database[7] and ARRS **GoldMiner**[8]. The former is a search engine indexing more than 80,000 documents and covering 150 manually selected health resources such as MEDLINE and the Cochrane Library. Its intent is to allow easy access to online evidence-based material for physicians [22]. The logs contain queries of 279,340 anonymous users from January 2011 to August 2012. GoldMiner consists of logs from an *image* search engine that provides access to more than 300,000 radiology images based on text queries of text associated with the images. Although the usage of an image search engine is slightly different from document search, previous work in the literature [30, 15] showed that the user search behaviour is similar.

Table A.8 summarizes the main statistics for the 5 datasets created and presents the aggregate numbers for laypeople and professional data.

[3] Obtained from http://www.gregsadetsky.com/aol-data/
[4] http://www.hon.ch/HONsearch/Patients/index.html
[5] dmoz.org
[6] The Latin 1 covers the majority of European languages, however it excludes the majority of Asian languages
[7] http://www.tripdatabase.com/
[8] http://goldminer.arrs.org

Preprocessing log files The first challenge dealing with different sources of logs is to normalize them. Hence, only the intersection of all possible fields was used: (1) timestamps, (2) anonymous user identification, and (3) keywords.

The package PunktWordTokenizer from NLTK (version 2.0.4)[9] was used to tokenize the query keywords. It isolates the punctuation tokens in a sentence, e.g., the query *(area: "Primary Care") women's health* was transformed into the list of tokens: ['(', 'area', ':', "'", 'Primary', 'Care', "'", ')', 'women', ''s', 'health']. This allows us to analyze punctuation tokens, often ignored in query log analysis. Stop words were not excluded nor was stemming used.

Sessions were defined as follows. They begin with a query and continue with the subsequent queries from the same user until a period of inactivity of over 30 minutes is found. This approach for sessions, as well as the 30-minutes threshold, is widely used in the literature [7, 31, 17]. We excluded extremely prolific users (over 100 queries in a single session), since they could represent "bots" rather than individuals.

All user queries were mapped to MeSH concepts (2011 edition) and Semantic Types [24] using NLM's Metamap (version 2011_v2) with default processing options [1].

After preprocessing, each one of the queries was converted into the following format: $\{timestamp, userId, query, meshIds, semanticTypes\}$, where the *timestamp*, *userId* and *query* are self explaining fields, *meshIds* and *semanticTypes* are the set of MeSH identifiers and semantic types generated by Metamap. These two concepts are detailed later in Section 3.1.2.

3.1.2 Analysis

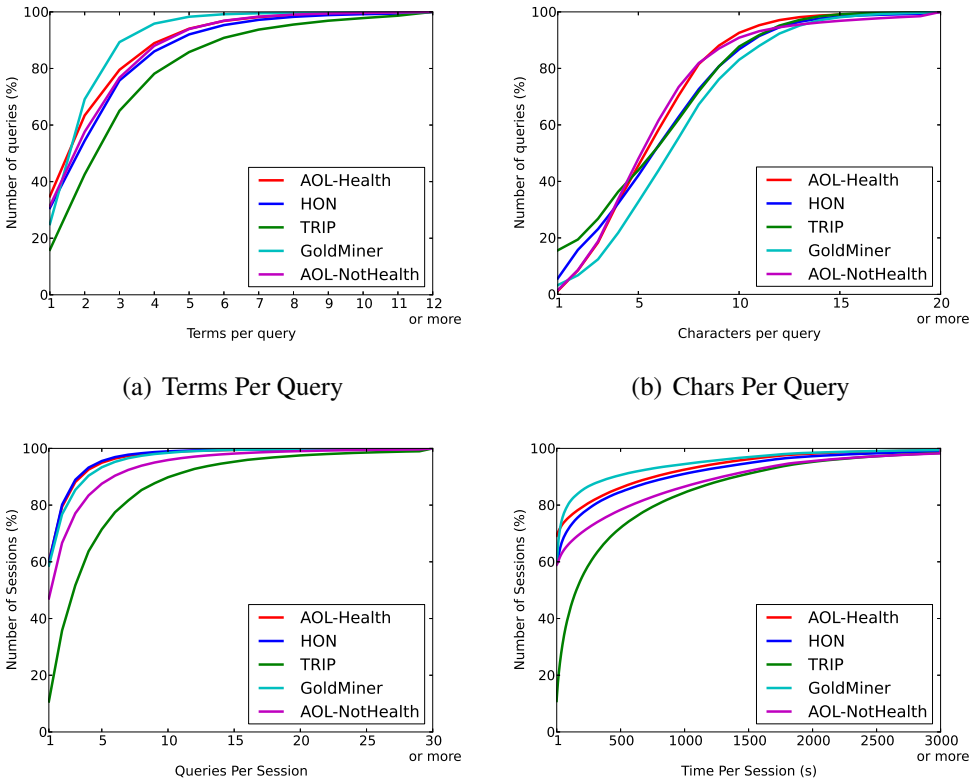
General Statistics Table A.8 shows some of the most important metrics used in the literature to characterize user iterations and Figure 3 shows the cumulative distribution function for some of the metrics shown in Table A.8. Together, the table and figure can depict the most important information about the datasets used.

Usually, the number of terms/characters per query, the number of queries per session and the time spent by users during a session may indicate how difficult a task is [8]. Here, we show that the number of terms per query is very similar for the laypeople dataset (HON and AOL-Health). Although the mean and median number of terms per query is higher for the TRIP logs, this is not the case for the GoldMiner logs, both belonging to the professional dataset. A similar behaviour happens for the time per session, where TRIP has a higher mean time in comparison with the other datasets.

In addition to the most common metrics, also in Table A.8, we analyse three other metrics: the number of queries: (1) expressing natural language, (2) containing medical abbreviations and (3) using Boolean operators.

The natural language queries are those in which the users express their needs through a complete sentence, e.g., *"how can bulimia be prevented"*. To capture this behaviour, a list of terms containing words like "what", "where", "how" and auxiliary verbs ("can", "would", etc.) was employed. The results show that the number of natural language queries is low, less than 3% of all queries, but the usage by laypeople is around 10 times more frequent than by professionals.

[9] http://nltk.org/api/nltk.tokenize.html



(a) Terms Per Query

(b) Chars Per Query

(c) Queries Per Session

(d) Time Per Session

Figure 3: Cumulative distribution frequency of general metrics

To find medical acronyms, we used lists of medical acronyms found on Wikipedia[10]. However, some abbreviations had to be manually removed, due to the fact that they were common words in English. For instance, TEN (toxic epidermal necrolysis) and several US states, such as FL (femur length) and CA (cancer, calcium, carcinoma). It is important to highlight that CAT (computed axial tomography) is one interesting example of a common word in both AOL and GoldMiner logs that was removed.

Many other dubious cases may have happened, leaving space for refinements. Nevertheless, it is possible to notice the more frequent use of medical acronyms in the health related logs (around 7%) than in the AOL-NotHealth log, showing that this can be an interesting feature to use when the intent is to find medical queries.

The use of Boolean operators was investigated in other medical work [22, 14] because it is widely used in PubMed, which has 21.8% of queries containing “and”, “or” or “not”. Meats et al. studied the TRIP database in 2004, finding 12% Boolean queries. We report a slight increase in the usage of Boolean queries for TRIP, 14%, and very low use in general for the other logs, less than 4%. We hypothesize that TRIP users are accustomed to PubMed, which explains this behaviour.

Finally, we report that we evaluated more than half a million health searchers, issuing more

[10] http://en.wikipedia.org/wiki/Acronyms_in_healthcare and http://en.wikipedia.org/wiki/List_of_medical_abbreviations, for example

than three million queries in more than one million sessions. Also, the proportion of health-related queries submitted to AOL is 2.7%, slightly smaller than the 3.4% found in [31] for the data collected in 2011 in a browser toolbar. This number confirms the trend of user migration from general to specialist search engines (it was 9.5% in 1997, 7.8% in 1999 and 7.5% in 2001 for Excite[29]).

Terms and Queries We depict the most popular queries and terms (only now excluding the stop words) used in all datasets, as well as their frequency among the queries in Table A.9. As was expected, AOL-NotHealth contains navigational queries and diverse terms related to entertainment. Similarly, some of the most popular queries in AOL-Health are navigational, with webmd.com appearing 3 times. The analysis of AOL-Health terms shows common health related concepts, with people searching information about hospitals in almost 2% of the cases. The queries in TRIP are similar to those in AOL-Health, however the terms used in TRIP reveal that users employed advanced searches with higher frequency. *Area* and *title*, for example are properties used like in PubMed, for instance in the query: *palliative care (area:oncology)*. The HON logs show that people constantly search for reliable content, frequently looking for URLs to check if they are trustable: it explains the high amount of punctuation (:&#) and the "http" among the most popular terms. Also, the second and fourth most popular query in HON are written in French and Spanish, showing the users offten want trustworthy sites in their own language. Finally, we clearly see the increase in difficulty and specificity of the most popular queries and terms found in the radiology query logs of GoldMiner.

MeSH Analysis An interesting measure calculated by Herskovic et al. [14] is the categorization of queries into the Medical Subject Headings (**MeSH**). MeSH is a controlled hierarchical vocabulary used by the National Library of Medicine in the USA for indexing journal articles in the life sciences field. The whole hierarchy contains more than 25 thousand of subject headings, with the most recent version containing 16 top categories such as "Anatomy" and "Diseases".

The MeSH analysis allows us to identify what are the most popular topics in the medical area according to the logs. We used MetaMap to map each query into one or more MeSH terms. For example, the query *lung cancer* is mapped to the MeSH terms *C04.588.894.797.520*, *C08.785.520* and *C08.381.540* all in the topmost *Disease* category (represented by the starting letter 'C'). As made in Herskovic analysis, the sum of scores for each query was always one. Therefore, if a query is mapped to four MeSH terms, each terms is weighted by 1/4. In the example discussed, all the terms belong to the same topmost category C (*Diseases*), weighting 1 to this category. Going further, the first two belong to the subcategory C08 (*Respiratory Disease*) and the last one to C04 (*Neoplasms Diseases*), weighting 2/3 to C08 and 1/3 to C04.

General statistics calculated for the mapping of queries in MeSH terms are shown in Table A.10. We present in Table A.11 the most popular categories for the first level of MeSH hierarchy, as well as we drilled down into the "Disease" category, in the second level of the hierarchy, to determine the most popular clinical topics. We also show the results obtained by Herskovic et al. for PubMed [14], in order to compare our findings. For sake of space, we limited to show only the categories in the first (second) hierarchical level which have more than 10% (5%) of the queries containing MeSH terms mapped to it. Exemplifying, to make the analysis clear, the 24.78% of queries for the *Neoplams* category in the GoldMiner logs, represents that 24.78% of 58.62% (for the *Diseases*) of all queries containing a MeSH term had some word

related to neoplasm, such as cancer or tumor. Finally, we sort the tables by the results found in PubMed.

Differently from PubMed, we found that the users are more interested in the diseases category, mainly the medical professional users, then followed by chemical and drugs. Nevertheless, the results for GoldMiner shows another trend for the second most popular category, focused on anatomy rather than on drugs, likely because a radiologist needs to identify the body parts present in the image to obtain more precise results.

The lower half of Table A.11, shows the second level analyses, specialized in the Diseases category. We confirmed that signs and symptoms is the most popular topic among the searches for all medical logs, but GoldMiner, where most often users search for neoplasms (e.g. cancer). Finally, the relatively high number of mappings in the **AOL-NotHeath**, Table A.10, is mapped to the categories "Geographic Location" and "Information Science", not directly related to health.

Sessions A series of queries, part of an information seeking activity, is defined as a session. We consider that, after issuing the first query, a user may act in four different ways: (1) repeat exactly the same query, (2) repeat the query adding one or more terms to increase precision, (3) reduce the number of terms to increase recall, or (4) reformulate the query changing some of the terms used. We ignore the first possibility because we cannot be sure if a user is really repeating the same query or just changing the result page, as some search engines record the same query as a result of a page change.

Table A.12 depicts the changes made by the users during the sessions. If during one single session a user adds a term to the previous query and then changes one word, we count one action in the row Exp.Ref (for expansion and reformulation). At the end, we divide the number of actions of each row by the total actions in the query log. Hence, Table A.12 shows that the most frequent user action is the reformulation alone but it is more likely to happen search engines targeting non-professionals, e.g., 78.47% of sessions in the AOL-Health logs had only reformulations. The percentage of reduction is relatively high in the GoldMiner logs due to the very frequent action of removing "mri" (for magnetic resonance imaging) and "ct" (computed tomography) from the original query. Also, it is evident that professional users are more persistent, as the last row of Table A.12 shows. More than 10% of the sessions in the professional search engines are composed of every type of action, while in AOL-Health this number is smaller than 1%.

Semantic Focus In this section, we attribute meaning to the users' queries in order to better understand their behaviour in a medical search context. We decide to use the same classes defined in Cartright, White and Horvitz [7]: Symptom, Cause and Remedy, so a direct comparison can be performed. A difference between their method and ours is that we classify the queries into the semantic classes using the Metamap tool by the US National Library of Medicine (NLM), instead of logistic regression classifier.

Metamap is responsible to map the queries to the Unified Medical Language System (UMLS) Metathesaurus, attributing semantic meaning for each query. A complete list of all 133 semantic types can be found online [24], but for completeness we link the abbreviations and their meaning in Table A.14. Also, in Table A.13, the most common semantic types and their frequency among the queries are shown. As we expect, the top 5 types in AOL-NotHealth are not really

related to the health domain, while the top 5 of all other datasets represent very well the logs. For example, the second most common semantic type for GoldMiner is related to parts of the body, as one might imagine for radiologic queries.

After a meticulous analysis of the semantic meaning assigned for the queries, we created the following classification (two examples of queries classified for each type are given for a better understanding):

- **Symptom:** sosy (cough; sore), lbtr (ph; high beta HCG), fndg (testicular cyst, stress)
- **Cause:** dsyn (diabetes; anemia), mobd (addiction; bipolar disorder), neop (lung cancer; tumor), patf (kidney stones; anaphylaxis)
- **Remedy:** clnd (gatorade; cough syrup), antb (antibiotic; penicillin), aapp(vectibix; degarilex), phsu (tylenol; mietamizol), imft (vaccine; acc antibody), vita (vitamin B12; quercetin)

We analyse the most popular semantic types found in the queries and show them in Table A.15, together with a direct comparison to Cartright et al. [7]. The first line shows the percentage of sessions where no semantic foci were found. The extremely high number found for AOL-NotHealth contrasts to the very small number found in the work of Cartright. The main reason for Cartright's result is linked to the way they created their dataset: *keeping only sessions that had at least one query containing a term in a wordlist extracted from a **list of symptoms** form the Merck medical dictionary*. Their preprocessing step also explains the fact that most of the sessions were concentrated only on searching for symptoms. Conversely, our analysis shows that the most common user focus is on causes rather than on symptoms. Also, the second most common focus is on a way to cure a disease. Once more, GoldMiner present a different behaviour, we hypothesize that the low number of sessions on remedies is explained by the fact that radiologists are not interested in remedies when searching for images. The 4.8% found for remedies and 4.7% for symptoms in the AOL-NotHealth logs may roughly represents the noise in our method to divide the AOL logs.

It is interesting to note that the difference between sessions only on symptoms and only on causes is small for the laypeople, but big for the medical professionals. It shows that in a clinical environment the physicians search for diseases rather than for the patient's symptoms. The laypeople, in turn, seek almost evenly for the symptoms, diseases and cures.

In Table A.16, we show the behaviour modifications along a session. One oscillation is characterized by a transition from one focus type to another and then back to the original type. Originally, this study was made to support the hypothetico-deductive searching process in that a user searches for a symptom, then a cause, and then returns to symptom [7]. However, in our analysis we found that the hypothetico-deductive process also often occurs when the users are looking for a remedy to cure a disease.

3.2 Classifying users

We introduce here one of the possible applications for the analysis made in the previous section, a classifier to infer users' expertise, separating laypeople from medical professionals.

As described in Section 2.3, a typical problem that affects laypeople, but usually not medical professionals, is the readability of documents [10, 9]. Often documents on the Internet are too

difficult for a regular user, but not for an expert user. It may affect the user's understanding, endangering his/her health integrity.

We defined a set of fifteen features to be used to classify users according to their expertise: laypeople or medical professionals. The features are listed in Table A.17 and are all defined using the full history of queries for each user. For example, *number of sessions* is the total number of sessions the user has started in the whole query log and *mean depth of MeSH terms* is defined as:

$$\text{MeanDepthOfMeSHTerms (u)} = \frac{\sum_{m \in \text{MeSHTerms(u)}} \text{depth(m)}}{|\text{MeSHTerms(u)}|}$$

where *MeSHTerms(u)* are all MeSH terms that MetaMap could find in the queries of a user u and *depth(m)* is a function that returns the level of depth of a given MeSH term m. All the features were normalized using the euclidean distance (l^2 normalization) and the Boolean features start with the words *is* or *has*.

We kept only users with more than 5 and less than 100 queries in the whole query log. Hence, we simulate an environment where the information about the users is not abundant nor scarce. We believe that it can make our simulation more real and challenging.

In total, 60,162 users were used as laypeople (merged from HON and AOL-Health datasets, the negative examples) and 133,702 as medical professionals (from TRIP and GoldMiner, the positive examples). Each single user is considered an example for the classifiers. Therefore in a 90/10 split, for example, we train a model using the logs of 54,145 users and test the model classifying each of other 6,017 users.

To measure the performance of each classifier, we used F_1 and **accuracy** scores, both highly used metrics for classification tasks. In special, F_1 is suitable to evaluate unbalanced distribution of examples as happens in this experiment.

Various classifiers from the Python package scikit-learn [27] were employed, using the default parameters for all of them. We considered as the baseline a classifier that assigns all the users to the majority class. We ran all the experiments using 10-fold cross-validation and comparisons to the baseline were made using a two tailed student's t-test with 99% confidence interval.

The results are summarized in Table A.18, as well as the result of t-test, with the symbol ▲ representing a statistically significant gain. The accuracy baseline shows the data distribution, while the F_1 baseline shows the F_1 score for the medical professional classes only, because it is the biggest class, thus it is a bigger challenge to obtain gains. Our results show that we had statistically significant gains for all classifiers, from 1.91% (F_1 for Naive Bayes (NB)) to 26.98% (accuracy for Random Forest (RF)). The K-Nearest Neighbors (K-NN - default K = 5) had results very close to Random Forest, while the Support Vector Machine (SVM) had intermediate results. We believe that tuning the parameters of the classifiers may improve even more the results, especially for SVM.

Table A.17 also reports the Gini importance score for each feature, generated by the Random Forest [4] classifier. Although, the other classifiers do not necessarily attribute the same importance for the features, it is a fair way to evaluate the feature significance. This normalized value is higher when the feature is more important, indicating how often a particular feature was selected for a split in a Random Forest, and how large its overall discriminative value was for the classification problem under study [23]. The number of sessions the user started was

the most important feature, followed by the number of queries issued and the mean time per session. As shown in Section 3.1.2, all three general metrics had different distributions and they Random Forest algorithm could use this knowledge in its favour. We leave as future work an in-depth analysis of how each feature affects the classification results, as the preprocessing steps (normalization, scale, data "binarization", etc.).

3.3 Discussion

We presented the analysis of four different query logs divided into five datasets. We sought patterns to understand how users search for health related information and if it is possible to differentiate a medical professional from a layperson.

Our analysis found that general metrics, such as the number of terms in queries or the number of queries in sessions, can be successfully applied as features to classify users according to their proficiency. We started the design of new metrics, such as the use of natural language and medical abbreviation. Although they were used in this work for the classification task, they were not the most important features. In the future, we are going to investigate how we can transform these metrics in more powerful classification features. It is also left as a future work the study of other methods to find medical abbreviations in queries rather than using a list of words, given that noisy and dubious acronyms are frequently found, and removing them may also be harmful.

It was discovered that laypeople very often only reformulate their queries, while professional users are more persistent: they are more likely to expand, to reduce and to reformulate their initial query during one single session.

We described two approaches to study the topics that users are most interested in: (1) using the topmost categories in MeSH hierarchy, (2) attributing a meaning to the semantic types that MetaMap use to classify queries. The former concluded that the topic "Disease" was the most popular, especially among the professional users. The latter confirmed this finding, conversely to previous work in the literature [7]. Cartright et al. performed a pioneering study on the user's medical intentions through query logs and we enhance it here. Besides the difference between their dataset and ours, the preprocessing steps applied were very distinct. An important future work is the investigation of new methods to filter health queries from general purpose query logs, as our method showed some error, such as the 4.8% for focus on remedies, and Cartright's method may have a bias in favor of symptoms. Newer widely used general query log, such as AOL once was, might also contribute to understanding modifications along the time on the medical user behaviour.

Finally, we plan to integrate the user classifier in the Khresmoi search engines to automatically predict user's expertise, providing the most suitable material according to the user's capability. We also plan to extend the classifier to assign users to a expertise index: 0 for a completely layperson and 100 for extremely expert user. Therefore, we could directly use the expertise index to re-rank the documents according to the user's readability.

4 Conclusion and Future Work

In this document we present the research performed and results obtained on document categorization, trustability and readability detection of health web pages, as well as on user classification.

We have also presented the integration of the document categorization and the document readability level within the "Khresmoi for Everyone" search engine. It has been shown in this research, that by tuning various parameters very good results can be achieved using various machine learning techniques in the domain of trustubility detection for health documents online.

The trustability work presented here will be further extended to other languages for which we are in possession of the training/test data. As for the readability research, it will be further extended to other training/test collections in English.

We also have shown a detailed log analysis, providing several features to classify users according to their expertise. In the future, we plan to add even more features and better understand how each feature contributes in our classifier. At the same time, we aim to investigate how to join the classifier to the ranking function, assing for each user the most relevant documents according to his/her medical litercy.

5 References

[1] A. R. Aronson. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. pages 17–21, 2001.

[2] Antoine Borst, Arnaud Gaudinat, Natalia Grabar, and Célia Boyer. Alexically-based distinction of readability levels of health documents. *Acta Informatica Medica*, 16(4):72–75, 2008.

[3] Célia Boyer, Vincent Baujard, and Antoine Geissbuhler. Evolution of Health Web certification through the HONcode experience. *Stud Health Technol Inform*, 169:53–7, 2011.

[4] Leo Breiman. Random forests. In *Machine Learning*, pages 5–32, 2001.

[5] David J. Brenes and Daniel Gayo-Avello. Stratified analysis of AOL query log. *Inf. Sci.*, 179(12):1844–1858, May 2009.

[6] Carrión, Fernández, and Toval. Are personal health records safe? a review of free web-accessible personal health record privacy policies. *Jmir*, 14(4), 2012.

[7] Marc-Allen Cartright, Ryen W. White, and Eric Horvitz. Intentions and attention in exploratory health search. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, SIGIR '11, pages 65–74, New York, NY, USA, 2011. ACM.

[8] Sergio Duarte Torres, Djoerd Hiemstra, and Pavel Serdyukov. Query log analysis in the context of information retrieval for children. In *Proceeding of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 847–848, New York, July 2010. ACM.

[9] Jean Anderson Eloy, Shawn Li, Khushabu Kasabwala, Nitin Agarwal, David R. Hansberry, Soly Baredes, and Michael Setzen. Readability assessment of patient education materials on major otolaryngology association websites. *Otolaryngology – Head and Neck Surgery*, 2012.

[10] Daniela B Friedman, Laurie Hoffman-Goetz, and Jose F Arocha. Readability of cancer information on the internet. *J Cancer Educ*, 19(2):117–22, 2004.

[11] A. Gaudinat, M. Joubert, S. Aymard, L. Falco, C. Boyer, and M. Fieschi. Wrapin: new generation health serach engine using umls knowledge sources for mesh term extraction from health documentation. *Studies in health technology and informatics*, 107:356–360, 2004.

[12] Arnaud Gaudinat, Natalia Grabar, and Célia Boyer. Machine learning approach for automatic quality criteria detection of health web pages. In Klaus A. Kuhn, James R. Warren, and Tze-Yun Leong, editors, *MedInfo*, volume 129 of *Studies in Health Technology and Informatics*, pages 705–709. IOS Press, 2007.

[13] Allan Hanbury, William Belle, Nolan Lawson, Ljiljana Dolamic, Natalia Pletneva, and Matthias Samwald. D8.3: Prototype of a first search system for intensive tests. *Khresmoi project public deliverable*, 2012.

[14] Jorge R. Herskovic, Len Y. Tanaka, William R. Hersh, and Elmer V. Bernstam. A Day in the Life of PubMed: Analysis of a Typical Day's Query Log. *JAMIA*, 14(2):212–220, 2007.

[15] Vera Hollink, Theodora Tsikrika, and Arjen P. de Vries. Semantic search log analysis: A method and a study on professional image search. *Journal of the American Society for Information Science and Technology*, 62(4):691–713, 2011.

[16] B.J. Jansen, A. Spink, and I. Taksai. *Handbook of research on Web log analysis*. Information Science Reference - IGI Global Publishing, Hershey, PA, 2008.

[17] Rosie Jones and Kristina Lisa Klinkner. 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*, CIKM '08, pages 699–708, New York, NY, USA, 2008. ACM.

[18] Wong LM, Yan H, and al. 'urologists in cyberspace: A review of the quality of health information from american urologists' websites using three validated tools. *Can Urol Assoc J.*, 7(3-4):100–107, 2013.

[19] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to information retrieval*. Cambridge University Press, 2008.

[20] Paul McNamee and James Mayfield. Character n-gram tokenization for european language text retrieval. *Information Retrieval*, 7(1-2):73–97, 2004.

[21] Paul McNamee, Charles K. Nicholas, and James Mayfield. Don't have a stemmer?: be un+concern+ed. In Sung-Hyon Myaeng, Douglas W. Oard, Fabrizio Sebastiani, Tat-Seng Chua, and Mun-Kew Leong, editors, *SIGIR*, pages 813–814. ACM, 2008.

[22] E. Meats, J. Brassey, C. Heneghan, and P. Glasziou. Using the Turning Research Into Practice (TRIP) database: how do clinicians really search? *J Med Libr Assoc*, 95(2):156–63, 2007.

[23] Bjoern H. Menze, B. Michael Kelm, Ralf Masuch, Uwe Himmelreich, Peter Bachert, Wolfgang Petrich, and Fred A. Hamprecht. A comparison of random forest and its gini importance with standard chemometric methods for the feature selection and classification of spectral data. *BMC Bioinformatics*, 10, 2009.

[24] MetaMap - Semantic Type Mappings. http://metamap.nlm.nih.gov/SemanticTypeMappings 2011AA.txt, May 2013.

[25] Nada Naji, Jacques Savoy, and Ljiljana Dolamic. Recherche d'information dans un corpus bruité (ocr). In Gabriella Pasi and Patrice Bellot, editors, *CORIA*, pages 271–286. Éditions Universitaires d'Avignon, 2011.

[26] Greg Pass, Abdur Chowdhury, and Cayley Torgeson. A picture of search. In *Proceedings of the 1st international conference on Scalable information systems*, InfoScale '06, New York, NY, USA, 2006. ACM.

[27] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[28] C. J. VAN RIJSBERGEN. *Information Retrieval*. Butterworths, London, UK, 1979.

[29] A. Spink, Y. Yang, J. Jansen, P. Nykanen, D. P. Lorence, S. Ozmutlu, and H. C. Ozmutlu. A study of medical and health queries to web search engines. *Health Information & Libraries Journal*, 21(1):44–51, March 2004.

[30] Theodora Tsikrika, Henning Müller, and Charles E. Kahn Jr. Log analysis to understand medical professionals' image searching behaviour. In *Medical Informatics Europe*, 2012.

[31] Ryen W. White and Eric Horvitz. Studies of the onset and persistence of medical concerns in search logs. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '12, pages 265–274, New York, NY, USA, 2012. ACM.

[32] K. Williams and RA. Calvo. A framework for document categorization. *7th Australian document computing symposium*, 2002.

[33] Olena Zubaryeva and Jacques Savoy. Investigation in statistical language-independent approaches for opinion detection in english, chinese and japanese. In *CLIAWS3 '09*, pages 38–45, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

Appendices

A Tables

Table A.1: Comparison of different thresholds

Class	Threshold											
	1			3			6			10		
	P	R	F	P	R	F	P	R	F	P	R	F
Alcohol	0.69	0.79	0.73	1.00	0.79	**0.88**	1.00	0.79	**0.88**	1.00	0.71	0.83
Cancer	0.53	1.00	0.69	0.88	0.96	**0.92**	0.95	0.87	0.91	1.00	0.83	0.90
Cardiovascular disease	0.50	1.00	0.67	1.00	1.00	**1.00**	1.00	0.90	0.95	1.00	0.90	0.95
Drugs	0.79	0.83	**0.81**	1.00	0.67	0.80	1.00	0.44	0.62	1.00	0.28	0.43
Env. Health	0.58	0.96	0.73	0.84	0.88	**0.87**	0.89	0.80	0.85	0.89	0.76	0.82
Food safety	0.58	0.88	0.70	0.90	0.73	**0.81**	0.95	0.69	0.80	0.94	0.62	0.74
HIV/Aids	0.38	1.00	0.56	0.77	1.00	0.87	0.90	0.90	**0.90**	0.90	0.90	**0.90**
Influenza	0.56	0.91	0.69	0.77	0.91	0.83	0.82	0.82	0.82	0.90	0.82	**0.86**
Mental health	0.48	0.93	0.64	0.93	0.93	**0.93**	0.93	0.87	0.90	0.92	0.80	0.86
Nutrition	0.41	0.95	0.58	0.85	0.85	**0.85**	1.00	0.65	0.79	1.00	0.60	0.75
Rare diseases	0.62	0.89	0.73	1.00	0.78	**0.88**	1.00	0.78	0.87	1.00	0.78	**0.88**
Tobacco	0.56	0.93	0.70	0.87	0.93	**0.90**	0.93	0.87	0.90	0.93	0.87	**0.90**

Table A.2: Size of the learning corpora (in number of extracts)

Language	English	French	German	Spanish	Italian	Dutch
Authority	2812	2338	493	894	500	107
Complementarity	2835	2005	567	819	490	106
Privacy	2683	2055	470	850	33	100
Reference	2349	1888	405	707	358	96
Justifiability	872	827	35	310	166	37
Transparency	2861	2349	604	881	414	121
Financial disclosure	2700	2098	546	814	489	103
Advertising policy	1412	627	246	433	255	15
Date	2794	2158	570	862	505	109

Table A.3: Results for Authority, Complementarity and Privacy

Alg.	Token	Selec.	%Kept	Authority			Complementarity			Privacy		
				P	R	F	P	R	F	P	R	F
NB	W1	DF	0.3	0.50	**0.87**	0.63	0.83	0.95	0.89	0.70	0.98	0.82
NB	W1	DF	0.5	0.50	0.86	0.63	0.83	0.95	0.89	0.69	0.98	0.81
NB	W1	DF	0.8	0.50	0.85	0.63	0.84	0.95	0.89	0.69	0.98	0.81
NB	C4	DF	0.8	0.46	0.86	0.60	0.82	0.95	0.88	0.68	0.98	0.81
NB	COOC	DF	0.8	0.53	0.78	0.63	0.78	**0.96**	0.86	0.78	**0.99**	0.87
SVM	W1	DF	0.8	**0.70**	0.63	**0.66**	0.89	0.91	**0.90**	**0.96**	0.97	**0.96**
SVM	C4	DF	0.8	0.68	0.64	**0.66**	0.87	0.90	0.89	**0.96**	0.96	0.96
SVM	COOC	DF	0.8	0.67	0.53	0.59	**0.90**	0.89	0.89	**0.96**	0.94	0.95
NB	W1	ZS	0.8	0.52	0.82	0.64	0.82	0.94	0.88	0.67	0.96	0.79
SVM	W1	ZS	0.8	**0.70**	0.61	0.65	0.89	0.90	0.89	0.94	0.94	0.94

Table A.4: Results for References, Justifiability and Transparency

Alg.	Token	Selec.	%Kept	References			Justifiability			Transparency		
				P	R	F	P	R	F	P	R	F
NB	W1	DF	0.3	0.40	0.81	0.54	0.40	0.65	0.50	0.85	**0.97**	0.90
NB	W1	DF	0.5	0.41	0.79	0.54	0.45	0.59	0.51	0.85	0.96	0.90
NB	W1	DF	0.8	0.43	0.77	0.55	0.49	0.50	0.50	0.86	0.95	0.90
NB	C4	DF	0.8	0.37	**0.82**	0.51	0.37	**0.67**	0.47	0.80	0.95	0.87
NB	COOC	DF	0.8	0.49	0.60	0.54	0.47	0.37	0.41	0.81	0.86	0.83
SVM	W1	DF	0.8	**0.62**	0.61	**0.61**	**0.51**	0.50	0.51	**0.94**	0.95	**0.95**
SVM	C4	DF	0.8	0.58	0.60	0.59	0.49	0.56	**0.52**	0.93	0.94	0.94
SVM	COOC	DF	0.8	0.50	0.38	0.43	0.41	0.29	0.34	0.85	0.80	0.83
NB	W1	ZS	0.8	0.44	0.74	0.55	**0.51**	0.44	0.47	0.87	0.93	0.90
SVM	W1	ZS	0.8	0.59	0.58	0.59	**0.51**	0.46	0.48	0.93	0.94	0.94

Table A.5: Results for Financial Disclosure, Advertising and Data

Alg.	Token	Selec.	%Kept	Financial			Advertising			Date		
				P	R	F	P	R	F	P	R	F
NB	W1	DF	0.3	0.56	**0.94**	0.70	0.51	**0.94**	0.66	0.94	**0.95**	**0.95**
NB	W1	DF	0.5	0.56	**0.94**	0.70	0.54	0.91	0.68	0.94	**0.95**	0.94
NB	W1	DF	0.8	0.57	0.92	0.71	0.60	0.85	0.70	0.94	0.94	0.94
NB	C4	DF	0.8	0.53	0.92	0.67	0.53	**0.94**	0.67	0.92	**0.95**	0.93
NB	COOC	DF	0.8	0.59	0.85	0.70	0.61	0.70	0.65	0.85	0.88	0.86
SVM	W1	DF	0.8	**0.79**	0.79	**0.79**	**0.74**	0.81	**0.78**	**0.95**	**0.95**	**0.95**
SVM	C4	DF	0.8	0.77	0.76	0.77	0.72	0.81	0.76	0.94	**0.95**	0.94
SVM	COOC	DF	0.8	0.78	0.67	0.72	0.64	0.72	0.68	0.92	0.82	0.87
NB	W1	ZS	0.8	0.58	0.89	0.70	0.64	0.80	0.71	0.93	0.90	0.91
SVM	W1	ZS	0.8	**0.79**	0.79	**0.79**	**0.74**	0.79	0.77	0.94	0.92	0.93

Table A.6: Readability English lexical

Threshold	Precision	Recall	F-measure
17.63	0.8137	0.5297	0.6383
22.00	0.8426	0.5581	0.6714
26.2	0.8663	0.7177	0.7850
28.00	0.8721	0.7608	0.8127
30.00	0.8997	0.8289	0.8628
32.00	0.9094	0.8817	0.8953
33.00	**0.9136**	0.8965	0.9050
34.00	0.9100	0.9050	**0.9075**
35.00	0.8988	**0.9094**	0.9040
37.00	0.8414	0.8731	0.8569
40.00	0.7899	0.8183	0.8038

Table A.7: Readability French learning

Algorithm	Tokenizer	Feature-Selector	Features-Kept	Precision	Recall	F-measure
NB	W1	DF	all	0.9750	0.9762	0.9753
NB	W1	DF	0.8	0.9752	0.9753	0.9749
NB	W1	DF	0.5	0.9751	0.9752	0.9747
NB	COOC	DF	0.8	0.9827	0.9844	0.9834
NB	C4	DF	0.8	0.9707	0.9707	0.9702
SVM	W1	DF	0.8	0.9699	0.9691	0.9693
SVM	COOC	DF	0.8	**0.9911**	**0.9909**	**0.9910**
SVM	C4	DF	0.8	0.9484	0.9596	0.9483
NB	W1	ZS	0.8	0.9747	0.9744	0.9739
SVM	W1	ZS	0.8	0.9769	0.9765	0.9765

Table A.8: General Statistics

Dataset	AOL-Health	HON	TRIP	GoldMiner	AOL-NotHealth
Total number of users	165,269	68,952	279,340	45,090	655,398
Total number of queries	930,862	235,844	1,798,072	219,407	34,159,571
Mean terms Per Query	2.47 (±1.75)	2.75 (±2.46)	3.40 (±2.33)	2.28 (±2.54)	2.46 (±1.87)
Median terms Per Query	2.00	2.00	3.00	2.00	2.00
Total number of Sessions	453,751	117,632	344,038	100,843	10,586,349
Mean Queries Per Session	2.05 (±2.65)	2.00 (±2.63)	5.20 (±5.95)	2.18 (±2.57)	3.23 (±4.60)
Median Queries Per Session	1.00	1.00	3.00	1.00	2.00
Mean Time Per Session (sec)	209 (±537)	257 (±649)	471 (±758)	163 (±520)	382 (±805)
Median Time Per Session (sec)	0	0	155	0	0
% Queries with Natural Language	2.95	1.29	0.27	0.14	1.55
% Queries with Medical Acronym	5.25	8.17	7.05	7.18	2.05
% Queries using Booleans	3.34	2.57	14.06	1.92	1.93

Aggregate Data	Laypeople	Professional	Non-medical
Total number of users	234,221	324,430	655,398
Total number of Queries	1,166,706	2,017,479	34,159,571
Total number of Sessions	571,383	444,881	10,586,349

Table A.9: Top queries and terms and their relative frequency (%) among all queries

Rank.	AOL-Health		HON		TRIP		GoldMiner		AOL-NotHealth	
	String	Freq	String	Freq	String	Freq	String	Freq	String	Freq
QUERIES										
1	webm	0.66	trustworthy health sites	1.91	skin	0.29	mega cisterna magna	0.44	google	0.96
2	web md	0.34	sites de confiance	1.38	diabetes	0.22	baastrup disease	0.40	ebay	0.42
3	weight watchers	0.21	diabetes	0.45	asthma	0.17	toxic	0.23	yahoo	0.30
4	pregnancy	0.17	cancer	0.23	hypertension	0.14	limbus vertebra	0.22	mapquest	0.26
5	herpes	0.15	HONcode sitios	0.22	stroke	0.13	cystitis cystica	0.20	google.com	0.23
6	diabetes	0.13	webmd	0.20	osteoporosis	0.11	thornwaldt cyst	0.14	myspace.com	0.22
7	shingles	0.12	sleep apnea syndromes	0.17	area: pediatrics	0.11	buford complex	0.13	myspace	0.22
8	lane bryant	0.12	asthma	0.15	low back pain	0.10	hemangioma splenic	0.13	yahoo.com	0.17
9	lupus	0.12	diabete	0.13	copd	0.10	throckmorton sign	0.12	www.yahoo.com	0.13
10	webmd.com	0.11	alzheimer	0.12	breast cancer	0.09	double duct sign	0.12	www.google.com	0.13
TERMS										
1	hospital	1.73	sites	3.38	(	12.73	cyst	3.17	free	1.27
2	cancer	1.61	health	2.87	)	12.69	mri	1.89	google	1.05
3	medical	1.14	”	2.20	“	12.50	disease	1.80	http	0.80
4	hair	1.13	trustworthy	1.94	:	9.02	ct	1.75	county	0.67
5	pain	1.11	cancer	1.87	area	3.27	fracture	1.68	pictures	0.63
6	disease	1.06	:	1.56	treatment	3.03	tumor	1.65	yahoo	0.55
7	symptoms	1.00	confiance	1.44	cancer	2.56	syndrome	1.47	lyrics	0.54
8	blood	0.97	http	1.25	title	2.48	liver	1.26	school	0.51
9	pregnancy	0.96	;	1.20	pain	2.13	pulmonary	1.22	com	0.51
10	health	0.92	&	1.13	care	2.10	bone	1.16	myspace	0.50
11	surgery	0.79	diabetes	1.09	children	1.98	renal	1.13	ebay	0.48
12	weight	0.78	#	1.08	therapy	1.81	sign	1.12	florida	0.46
13	breast	0.75	syndrome	0.76	diabetes	1.80	lung	1.11	sex	0.43
14	heart	0.74	,	0.74	disease	1.78	brain	1.08	sale	0.42
15	center	0.69	disease	0.73	pregnancy	1.70	cell	1.00	www	0.42

Table A.10: General MeSH Statistics

Metric	AOL-H	HON	TRIP	GM	AOL-NH
Queries containing MeSH terms (%)	69.60	50.73	80.71	63.04	47.41
MeSH terms per query	1.76	1.32	2.48	1.57	1.06
Disease terms per query	0.49	0.38	0.97	0.97	0.04

Table A.11: Queries by First level Category and by Disease Type according to MeSH Mappings

Category name	PubMed [14]	AOL-Health	HON	TRIP	GoldMiner	AOL-NotHealth
Popular MeSH Categories						
Chemicals and Drugs	24.61	11.07	14.57	12.57	2.23	4.71
Diseases	20.16	24.51	28.25	**39.15**	**58.62**	2.98
Biological Sciences	10.79	6.74	6.18	4.82	2.45	5.52
Anatomy	10.27	9.35	8.62	4.71	24.99	3.34
Analytical, Diagnostic and Therapeutic Techniques and Equipment	7.42	6.92	7.10	12.42	6.14	2.92
Health Care	2.44	9.04	10.68	6.28	0.39	6.61
Information Science	2.08	5.02	4.13	0.66	0.30	21.4
Geographic Locations	0.85	3.36	1.91	1.73	0.38	14.16
Popular MeSH Diseases						
Pathological Conditions, signs and symptoms	13.03	17.72	15.96	13.77	13.57	29.07
Nervous System diseases	8.79	9.15	8.35	10.41	6.23	14.06
Neoplasms	7.99	7.99	12.08	5.67	**24.78**	6.05
Cardiovascular diseases	7.35	5.03	6.14	9.11	5.63	2.97
Immune system diseases	6.93	2.34	3.01	2.64	1.26	1.74
Bacterial infections and mycoses	5.30	5.19	4.45	4.78	3.41	2.25
Nutritional and metabolic diseases	5.19	2.80	5.65	4.84	1.35	2.24
Skin and connective tissue diseases	4.80	7.35	5.20	4.93	1.96	4.90
Musculoskeletal diseases	4.77	5.17	4.22	4.64	7.57	5.52
Virus diseases	4.27	5.16	4.73	2.20	0.65	3.25
Digestive system diseases	4.08	5.33	4.63	6.02	7.32	1.64
Female genital diseases and pregnancy complications	1.49	4.74	2.47	5.22	2.92	3.90

Table A.12: Modifications along the sessions

Dataset	AOL-H	HON	TRIP	GM	AOL-NH
Expansion	10.56	13.21	14.85	5.96	3.73
Reduction	1.89	2.48	4.35	**9.61**	0.84
Reformulation	**78.47**	**59.65**	43.95	49.57	80.36
Exp.Red.	0.92	1.61	5.09	3.54	0.57
Exp.Ref.	6.11	15.15	15.27	8.28	9.59
Red.Ref.	1.19	3.16	5.63	12.01	2.08
Exp.Red.Ref.	0.87	4.73	**10.85**	**11.04**	2.83

Table A.13: Top 5 semantic types used

Rank	AOL-Health		HON		TRIP		GoldMiner		AOL-NotHealth	
	Type	Freq(%)	Type	Freq(%)	Type	Freq(%)	Type	Freq(%)	Type	Freq(%)
1	dsyn	14.28	dsyn	13.22	**dsyn**	**31.58**	**dsyn**	**29.09**	mnob	25.33
2	phsu	12.29	phsu	10.68	topp	17.80	bpoc	21.75	geoa	12.79
3	mnob	12.14	fndg	8.08	phsu	16.44	neop	15.21	inpr	11.84
4	fndg	10.11	ftcn	7.83	fndg	14.55	fndg	9.50	qlco	9.54
5	ftcn	9.81	gngm	7.65	ftcn	14.44	qlco	6.96	ftcn	7.35

Table A.14: Semantic types used and their meaning

Abbreviation	Meaning
aapp	Amino Acid, Peptide, or Protein
antb	Antibiotic
bpoc	Body Part, Organ, or Organ Component
clnd	Clinical Drug
dsyn	Disease or Syndromee
fndg	Finding
ftcn	Functional Concept
geoa	Geographic Area
gngm	Gene or Genome
imft	Immunologic Factor - vaccine, e.g.
inpr	Intellectual Product
lbtr	Laboratory or Test Result
mnob	Manufactured Object
mobd	Mental or Behavioral Dysfunction
neop	Neoplastic Process
patf	Pathologic Function
phsu	Pharmacologic Substance
qlco	Qualitative Concept
sosy	Sign or Symptom
topp	Therapeutic or Preventive Procedure
vita	Vitamin

Table A.15: Semantic Focus

Dataset	AOL-Health	HON	TRIP	GoldMiner	AOL-NotHealth	Cartright et al.
Nothing	58.6	59.7	26.8	32.2	**86.2**	**3.9**
Symptom	8.4	5.5	8.6	7.9	4.7	**63.8**
Cause	**14.4**	**15.0**	**30.0**	**50.6**	2.8	5.3
Remedy	11.9	9.8	10.5	**1.6**	**4.8**	1.1
Symptom and Cause	4.0	4.9	10.2	6.1	0.5	22.6
Symptom and Remedy	0.9	0.9	2.3	0.2	0.5	2.0
Cause and Remedy	1.3	3.0	8.1	1.1	0.4	0.4
All three	0.5	1.2	3.5	0.3	0.1	0.8

Table A.16: Cycle Sequence

Dataset	AOL-Health	HON	TRIP	GoldMiner	Cartright et al.
Sessions with oscillations (%)	6.0	7.6	**40.0**	5.4	16.2
Symptom→Cause→Symptom	**30.9**	**22.0**	**21.5**	**30.4**	**51.4**
Cause→Symptom→Cause	**31.2**	**25.6**	**23.4**	**45.3**	**38.4**
Symptom→Remedy→Symptom	8.2	8.3	8.2	2.3	5.1
Remedy→Symptom→Remedy	8.2	8.9	8.8	2.1	2.7
Cause→Remedy→Cause	**10.1**	**16.7**	**19.3**	**11.2**	1.5
Remedy→Cause→Remedy	**11.4**	**18.5**	**18.8**	**8.7**	0.9

Table A.17: Features ranked by the importance scores

Feature	Score
number of sessions	30.70
number of queries	10.30
mean time per session	9.54
mean terms per query	8.38
has reformulated queries	6.99
has searched for causes	6.71
mean depth of MeSH terms	6.53
has searched for non-medical content	4.23
has searched for remedies	3.29
has searched for symptons	3.24
is using natural language	3.06
is using medical abbreviation	2.24
has exp. red. and ref. queries	1.96
has expanded queries	1.63
has reduced queries	1.20

Table A.18: Classification results

Classifier	Accuracy	F-1
Baseline	68.97	81.65
NB	73.35 ± 0.41 (6.35% ▲)	83.20 ± 0.22 (1.91% ▲)
K-NN	87.02 ± 0.15 (26.18% ▲)	90.97 ± 0.09 (11.44% ▲)
RF	87.58 ± 0.29 (26.98.% ▲)	91.05 ± 0.20 (11.54% ▲)
SVM	80.42 ± 0.63 (16.60% ▲)	87.40 ± 0.86 (7.05% ▲)