Identification and Classification of Health Queries: Co-occurrences vs. Domain-Specific Terminologies

Carla Teixeira Lopes[1] and Cristina Ribeiro[1,2]

[1]DEI, Faculdade de Engenharia, Universidade do Porto
[2]INESC TEC

{ctl, mcr}@fe.up.pt

Abstract

Identifying the user's intent behind a query is a key challenge in Information Retrieval. This information may be used to contextualize the search and provide better search results to the user. The automatic identification of queries targeting a search for health information allows the implementation of retrieval strategies specifically focused on the health domain. In this paper, two kinds of automatic methods to identify and classify health queries based on domain-specific terminology are proposed. Besides evaluating these methods, we compare them with a method that is based on co-occurrence statistics of query terms with the word "health". Although the best overall result was achieved with a variant of the co-occurrence method, the method based on domain-specific frequencies that generates a continuous output outperformed most of the other methods. Moreover, this method also allows the association of queries to the semantic tree of the Unified Medical Language System and thereafter their classification into appropriate subcategories.

Keywords: *Health Queries, Information Filtering, Information Retrieval, Health Vocabularies.*

1 - Introduction

Today, the Web is a major source of information worldwide and the use of popular search engines to seek health information is commonly practiced by Internet users. In 2011, 80% of Internet users in the United States used the Web to search for health information (Fox, 2011). According to Eysenbach & Kohler (2003), over 12 million health queries were made per day in Google in 2003. To provide more focused support and better retrieval services to users searching for health information, there is a need to automatically identify health queries, that is, queries that are intended to retrieve health-related information and are motivated by the need to seek health knowledge.

The classification of queries is used frequently to distinguish and categorize them according to major topic or subsets. This classification can be manual. It may also involve the comparison of a query with databases of queries or it may require machine-learning processes. Another possibility is the use of controlled vocabularies, or thesaurus of terms, in areas where the quality of these structures can be trusted.

As most health queries contain terms that can be mapped onto standardized health/medical vocabularies (McCray, Loane, Browne, & Bangalore, 1999; Zeng et al., 2006), we propose two methods to detect consumer health queries that would leverage on existing high-quality health vocabularies. Considering the search results of Google and Yahoo! we have also replicated a method proposed by Eysenbach & Kohler (2003), a method that is based on the co-occurrences of query terms with the word "health" in web documents

The rest of this paper is organized as follows. Section 2 describes related work regarding topic detection and, more specifically, the identification of health queries. Section 3 summarizes the rationale of the co-occurrence method (COM) and the methodology we adopted for its replication in three variants. In Section 4, we propose two kinds of methods based on domain-specific semantic structures, that is, the structures we have used here. Section 5 presents the evaluation results while Section 6 discusses implications of our findings. We then conclude in Section 7.

2 – Related Work

A manual approach to classify web queries is straightforward. Usually several assessors are involved in the classification process; and, to reduce the subjectivity, more than one person typically is asked to classify the very same query. If and when a consensus is not found initially, either another element is added to ease the classification or a discussion between the adjudicators is promoted to reach a consensus. In a study that focused on studying queries that users submit to search engines, Amanda Spink, Wolfram, Jansen, and Saracevic (2001) manually classified a sample of 2,414 queries submitted to the Excite search Engine into 11 categories. Focusing on the study of health queries submitted to search engines, Spink et al. (2004) also do a manual classification of queries to select the ones related to the topic of health. Despite being a popular approach, manual classification is slow and represents a tedious process requiring the availability of one or more human classifiers. In some cases, the huge volume of queries may even make the classification task impracticable; for these reasons, automatic methods have been proposed.

In Information Retrieval (IR), several approaches to detect topics in documents and collections of documents have emerged. Some methods are based on mathematical models, for example, the method of Latent Semantic Analysis, which is a method based on co-occurrences of terms in the collection to reduce the semantic context of the documents (Landauer, Foltz, & Laham, 1998). Even so, as web queries are more or less short, these methods are not the most appropriate.

Another kind of methods involves the comparison of queries or terms with existing data structures; essentially, an attempt to find an exact match. The simplest approach takes the form of looking up the query in a set of manually classified queries. However, this is usually associated with a poor coverage and is highly dependent on the query stream dynamics (Beitzel et al., 2005). Term comparison with specific databases can improve coverage, but it requires additional processing like the tokenization of the queries. An example of this approach is discussed in Murata (2007), who automatically extracts news words from news websites and tries to find an exact match of one of these words with the query in order to detect breaking news from search queries.

An approach that naturally follows exact matching is "supervised learning", that is, training a classifier on the manually classified set of queries to detect features that could be useful in the classification of unlabeled queries. This is particularly challenging because web queries are typically short, thereby reducing the possible features to be used by the learner.

Finally, there also methods based on large datasets such as the Web itself. As noted, Eysenbach and Kohler (2003), whose work is specifically focused on the health domain, proposed a method to automatically classify search strings as health-related based on the proportion of pages on the Web containing the search string plus the word "health" and the number of pages containing only the search string.

Aside from Eysenbach and Kohler (2003), no other automatic mechanism to filter health queries was found in the current published literature. The nearest, but broader, topic is generic automatic query classification. An extensive state-of-the-art review on this topic is done by Beitzel et al. (2005). Still, given that our goal is restricted to the health domain, we believe there may be some simpler and more targeted strategies waiting to be developed.

3 – Co-Occurrence Method (COM)

The Eysenbach-Kohler (2003) co-occurrence method (COM) is based on the idea that health-related terms should co-occur more often with the word "health" than non-health terms. We named it COM because it uses co-occurrence statistics from web documents.

To replicate this method, for each query (Q) in the pool, two queries were submitted to a search engine: one (Q1) with the terms of the query Q and another (Q2) with the terms of Q plus the word "health". The health co-occurrence rate (cooc) of Q is calculated by the proportion of the total number of results of Q2 in the total number of results of Q1 as expressed in Equation 1, where is the set of terms that compose the query Q. If #results($terms_Q$) = 0, cooc(Q) = 0.

$$cooc(Q) = \frac{\#results(terms_Q \cup \{health\})}{\#results(terms_Q)} \quad (1)$$

This proportion is an indicator of the association degree of the query Q to the health domain because it represents the frequency of occurrence of Q's search terms and the word "health" in web pages.

Eysenbach and Kohler (2003) used Google in their study. Here, we will use both Google and Yahoo! to determine the number of results. For example, in Google, the query 'diabetes symptoms' has a health co-occurrence rate of 478,000/929,000=0.51 and the query 'Pavarotti' has a health co-occurrence rate of 359,000/6,440,000=0.06. The differences detected in the number of results of both search engines, also stated by Chitu (2007), made us combine the number of results returned by the two search engines in a third variant of this method. Therefore, we have implemented 3 variants with different health co-occurrence rates as expressed in Equations 2, 3 and 4.

$$\boldsymbol{G_{cooc_Q}} = \frac{\#\boldsymbol{google}(\boldsymbol{terms_Q} \cup \{\boldsymbol{health}\})}{\#\boldsymbol{google}(\boldsymbol{terms_Q})} \quad (2)$$

$$\boldsymbol{Y_{cooc_Q}} = \frac{\#\boldsymbol{yahoo}!(\boldsymbol{terms_Q} \cup \{\boldsymbol{health}\})}{\#\boldsymbol{yahoo}!(\boldsymbol{terms_Q})} \quad (3)$$

$$\boldsymbol{Y + G_{cooc_Q}} = \frac{\#\boldsymbol{google}(\boldsymbol{terms_Q} \cup \{\boldsymbol{health}\}) + \#\boldsymbol{yahoo}!(\boldsymbol{terms_Q} \cup \{\boldsymbol{health}\})}{\#\boldsymbol{google}(\boldsymbol{terms_Q}) + \#\boldsymbol{yahoo}!(\boldsymbol{terms_Q})} \quad (4)$$

As shown in Figure 1, we have developed scripts, one for each search engine, to automatically get the number of results returned for each query in Google and Yahoo! through each search engine's API. The script `classifyQueries.pl`, for each query, asks the script `numberofResults.pl` for the number of results of the query and the query plus the word "health". These values are then used to compute the health co-occurrence rate.

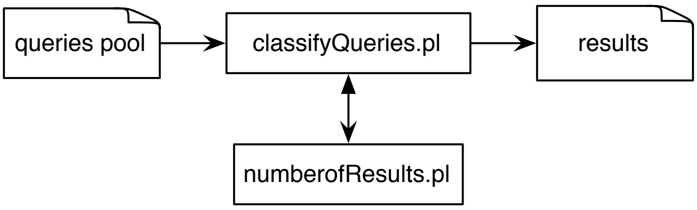

Figure 1 - COM global architecture - dataset files and Perl scripts

Following the computation of the health co-occurrence rate, this value was compared with several thresholds (0; 0.05; 0.1; 0.15; 0.2; ...; 0.95; 1). In each comparison, if the health co-occurrence rate was larger than or equal to the threshold, the query was considered to be a health-related query at that threshold.

4 – Methods based on Domain-Specific Terminologies

To take advantage of existing high-quality health vocabularies, we decided to propose two different methods using an existing health semantic structure, the Consumer Health Vocabulary (CHV). The first method has a binary output that can be either "health" (if the query has terms that are included in the CHV subset in use) or "non-health" as in all the other cases. We will denote the first method as CHV binary method. The second method computes a continuous output that quantifies the association degree of the query with the health domain and we will denote it as CHV continuous method. This latter method has been proposed in a previous work (Lopes, Dias, & Ribeiro, 2013)

In this section, we first describe the CHV and the Unified Medical Language System (UMLS), one of the most consistent and robust health semantic structures. Then, we will shift focus to the two CHV methods.

Health Semantic Structures

The Consumer Health Vocabulary connects "informal, common words and phrases about health to technical terms used by health care professionals" (Nlm, 2012). It is developed as an open source and collaborative initiative and can be used to improve IR systems, to help lay-people read and understand health-related information. CHV is part of the Unified Medical Language System (UMLS) since the 2011AA release and is also available from the CHV website (http://www.consumerhealthvocab.org) in file format or through an online browser. The latest version of CHV has 57,819 health concepts and 158,519 English concept strings. A CHV concept is identified by the UMLS unique identifier and may be associated with several synonymous strings to express that concept. Each CHV concept is also associated with a CHV preferred name and a UMLS preferred name. On the one hand, the CHV preferred name is the string that best represents that concept for health consumers and is defined by the CHV. On the other hand, the UMLS preferred name is the preferred string for that concept as defined by the UMLS.

The UMLS consists of three knowledge sources that can be used separately or together. The first is the Metathesaurus that has more than one million biomedical concepts from over 100 sources; next is the Semantic Network with 135 broad categories and 54 relationships between categories; and finally, there is the SPECIALIST Lexicon and Tools, which has lexical information and programs for language processing (Kleinsorge & Willis, 2007).

CHV Binary Method

As noted, this method produces a binary output. Given a query, this method either considers it is (or, is not) a health query. If the query has, at least, one term of the CHV subset in use, it falls in the former category. If it does not, it falls in the latter category.

11 variants of this method were tested, that is, 11 different subsets of the CHV: CHV1 (all terms), CHV2 (terms associated with the 200 most frequent concepts), CHV3 (terms associated with the 400 most frequent concepts), CHV4 (terms associated with the 600 most frequent concepts), CHV5 (terms associated with the 800 most frequent concepts), CHV6 (terms associated with the 1,000 most frequent concepts), CHV7 (terms existing in the UMLS preferred names), CHV8 (terms existing in the CHV preferred names), CHV9 (terms existing in the UMLS and CHV preferred names), CHV10 (6,000 most frequent terms) and CHV11 (10,000 most frequent terms). The list of most frequent terms was obtained from the CHV website.

As shown in Figure 2, we used two Perl scripts: `generateTermsList.pl` to generate a subset of health terms and `classifyQueries.pl`, similarly for all variants of the CHV binary method, which classifies queries. The `generateTermsList.pl` removes stop-words, using a list of stop-words provided by the University of Glasgow, and replaces special characters that may be misunderstood by regular expressions that are used later to parse the files. The `classifyQueries.pl` simply checks if any of the query terms is present in the terms list. If present, the query is classified as health-related.

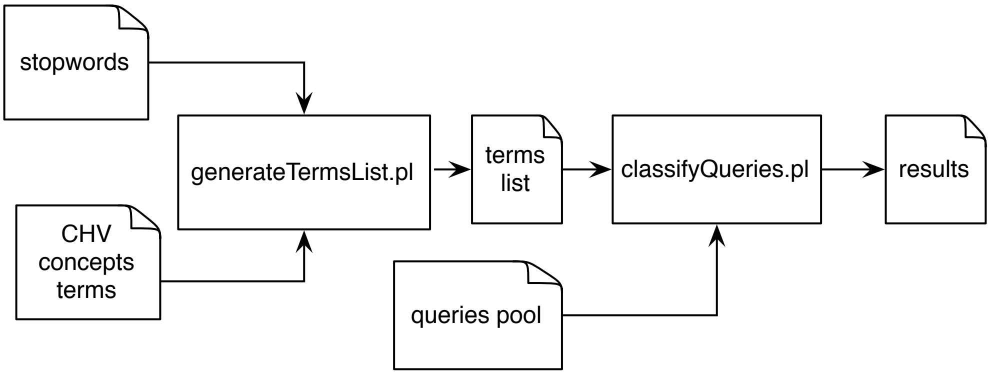

Figure 2 - CHV binary method global architecture - dataset files and Perl scripts

CHV Continuous Method

The variants of this method differ on the subset of the terms used to classify the queries. The presence of one term in a query is sufficient to classify it as a health query.

CHV Subsets

The CHV vocabulary contains concepts of several categories and some of them contain strings (e.g.: car, driving) that, when isolated from other health terms or concepts, are not useful to identify a health query. To avoid false positives, we obtain different subsets of the CHV vocabulary besides using the complete CHV. We defined four subsets: one with concept strings from UMLS categories containing concepts more likely to occur in consumer health queries (HEALTH), one with the consumer preferred string for each concept in the CHV (CHVP), one with the UMLS preferred string for each concept in the CHV (UMLSP) and the other with the MedlinePlus Health Topics source vocabulary concept strings (MEDP).

The HEALTH subset was created to include all the strings associated with concepts pertaining the UMLS semantic types that had a greater probability of embedding terms used by health consumers on their health searches. The semantic types containing mostly concepts related to the biology and chemistry fields were excluded, as their inclusion in health queries is unlikely. All the concepts falling directly under the following semantic types, whose numeration is the same as the one presented in the Semantic Network Browser, were included in the HEALTH subset:

```
A1.2 Anatomical Structure
A1.2.1 Embryonic Structure
A1.2.3 Fully Formed Anatomical Structure
A1.2.3.1 Body Part, Organ, or Organ Component
A1.2.3.2 Tissue

A1.4 Substance
A1.4.1.1.1 Pharmacologic Substance
A1.4.1.1.1.1 Antibiotic

A2.1.4.1 Body System

A2.1.5.2 Body Location or Region

A2.2 Finding
A2.2.2 Sign or Symptom

B1 Activity
B1.1 Behavior
B1.3.1 Health Care Activity
B1.3.1.2 Diagnostic Procedure
B1.3.1.3 Therapeutic or Preventive Procedure

B2.2.1 Biologic Function
B2.2.1.1.2 Organ or Tissue Function
B2.2.1.2 Pathologic Function
B2.2.1.2.1 Disease or Syndrome
B2.2.1.2.1.1 Mental or Behavioral Dysfunction
B2.2.1.2.1.2 Neoplastic Process
```

Auxiliary Structures

For each subset, we created an inverted index containing the unique terms mapped onto a list of unique identifiers for each concept string in the subset and their association degree with the concept string. The association degree of a term *t* to a concept string c, w_t^c, is computed as the ratio $tf_t^c/|c|$, where the numerator is the term frequency of *t i*n the concept string *c* and the denominator is the number of terms in concept string c. If we consider the CHV strings *tooth* and *dental infection*, the terms *dental* and *infection* would be associated with the second string with a probability of 0.5 and the term *tooth* with the first string with a probability of 1.

Combining Inverted Index Entries

In the classification process, queries are tokenized and, for each term, we retrieve the corresponding posting list from the inverted index. We then combine the posting lists of every term into a single list to which we call query list. As shown in Figure 3, two combination methods were tested. The first joins the lists and, when an identifier appears more than once, the w_t^c are added. The resulting list contains the weights of each CHV string in the query: w_c^q . This way we can easily identify if a query contains parts or entire health CHV strings. The second method (M2), joins the lists as M1, but also counts the occurrences of each CHV string in the query ($cf_{c,q}$). As a final step, we adjust the weights calculated in the first method as $w_c^q \times cf_{c,q}/|q|$, where $cf_{c,q}$ is the frequency of c in query q and $|q|$ the number of unique terms in q.

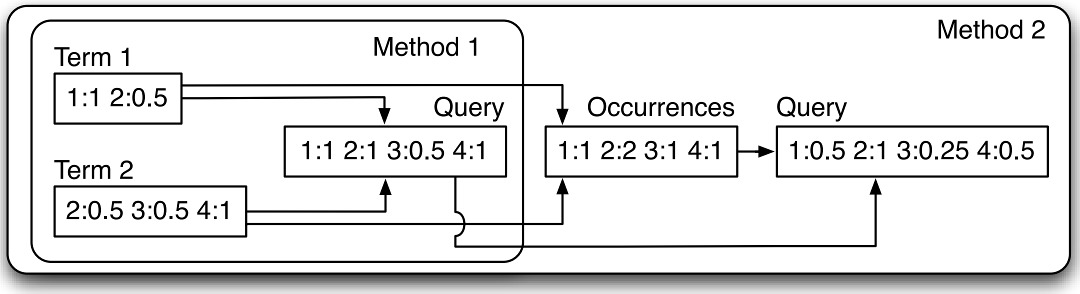

Figure 3 - Joining posting lists in Methods 1 and 2

Final Score Calculation

After obtaining the query list, we calculate the final score that will be used to classify the query as health related or not. To do this, we propose some variants for the two previous methods, as presented in Table 1. In that table, *Query* is the query list obtained after the combination of the terms' lists in each method, $tf_{h,q}$ is the number of terms in query q included in the inverted index, and $|q|$ is the number of unique terms in q. M1Max and M1MaxBoost use the maximum weight of the *Query* list under the assumption that, if a query is completely matched by a health concept, it is a health query. In M1Avg and M1AvgBoost we computed the average of the 5 largest probabilities in the query list.

Table 1 - Variants applied to the different methods

Variant	Formula	Boost
M1Max	$\max(\text{Query}) \times (tf_{h,q}/\lvert q\rvert)$	No
M1MaxBoost		Yes
M1Avg	$\text{avg}(top_1^5(\text{Query})) \times (tf_{h,q}/\lvert q\rvert)$	No
M1AvgBoost		Yes
M2Max	$\max(\text{Query})$	No
M2MaxBoost		Yes
M2Avg	$\text{avg}(\text{Query})$	No

The product used in the M1 variants lowers the score of the queries that have non-health terms even if the query matches an entire concept because a concept may change when a term is added. Consider, for example, the query "tooth piercing". As "tooth" is a CHV concept and the term "piercing" is not, without the final product the above query would score 1 instead of 0.5 as it is with the multiplication. This is not needed in the M2 variants because the M2 already uses the occurrences of each CHV concept string in the whole query.

To promote the queries that contain terms that appear more frequently in the CHV vocabulary, we decided to test the application of a boost value b to the term weights in a CHV string ($b \times w_t^c$). This boost is similar to the document frequency *df* used in IR and is equal to the number of strings in the CHV in which the term appears.

Classifying Health Queries

Queries that have the final score above a specific threshold will be classified as being health-related. We also used the UMLS semantic network to assign health categories to each query. For this purpose, we created an index similar to the one described above where terms are replaced by CHV strings and the posting lists contain categories and not strings. After obtaining the query list as explained above, we create another list with the category associated to each CHV string in the query list and the weight, w_c^q, previously associated with the string. If a category appears more than once, we select its maximum weight.

Vocabulary Translation

As one of the main disadvantages of a method based on vocabularies is its dependence on the language in which it was created, in the CHV continuous methods, we have tested if they can be applied with a Portuguese translated version of the CHV without much penalty on the results. To evaluate the efficacy of our method in Portuguese, we used the Google Translator API. We manually evaluated 1% of the total number of translated strings and concluded that 84.2% (95% CI: [82.3%, 85.9%]) of the translations were good, a very satisfactory outcome.

5 – Evaluation of the Methods

Methodology

To evaluate each method, we compared its classification with a classification done manually by a team of human assessors.

In the CHV binary method, the classification is immediately computed after the execution of the described scripts. In the methods with the continuous output, that is, the COM and the CHV continuous methods, the classification only occurs after the calculation of the cooc/final rate and its comparison with each threshold. The best thresholds are determined after the analysis of all collected data.

A collection of 20,000 web queries, randomly sampled from AOL Search in the Fall of 2004 was used in our method evaluation. Beitzel et al. (2005) used this collection in a research project where queries were classified into 20 topical categories by a team of approximately ten human assessors. One of these topical categories was health, where 1,197 queries were included. In the evaluation of the

COM and the CHV binary method, we used the 20,000 queries.

In the evaluation of the CHV continuous method, two datasets were used, one for each language. In Portuguese (PT), a collection of 1,522 queries manually classified by medical students was applied. The initial set of queries was composed of 1,553 queries extracted from the SAPO Saúde search engine and several assessors classified each query. When classification ties occurred, we excluded the query. In the final dataset, 55.6% of the queries were health queries. In order to obtain a dataset of a similar size for English, we used 1,647 queries from the AOL Search dataset where 1,197 are found to be health queries (72.7% of the entire sample).

For each method, measures such as sensitivity, specificity and accuracy were computed. These can be expressed in terms of probabilities of specific events: H_{hc} (query is classified as health-related in a human classification), NH_{hc} (query is classified as non-health-related in a human classification), H_{ac} (query is classified as health-related in an automatic classification) and NH_{ac} (query is classified as non-health-related in an automatic classification).

Sensitivity (SEN) is the number of true positives divided by the sum of true positives with false negatives. It can be expressed as the conditional probability of having an automatic classification of health-related, given that the query was classified as health-related by a human: $P(H_{ac}|H_{hc})$.

Specificity (SPC) is the number of true negatives divided by the sum of true negatives with false positives. It can be expressed as the conditional probability of having an automatic classification of non-health-related when the query was classified as non-health-related by a human: $P(NH_{ac}|NH_{hc})$.

Accuracy (ACC) is the tax of correct classifications (either as health-related or as non-health-related) and is expressed as stated in Equation 5.

$$\frac{P(H_{ac} \cap H_{hc}) + P(NH_{ac} \cap NH_{hc})}{P(H_{hc}) + P(NH_{hc})} \quad (5)$$

Besides computing these key measures, two Receiver Operating Characteristics (ROC) graphs for comparing the several discrete classifiers methods and the several continuous classifiers methods were also drawn. A ROC graph is a two-dimensional graph in which sensitivity is plotted on the Y-axis and the false positive rate (1-specificity) is plotted on the X-axis. It is a technique that depicts relative tradeoffs between benefits (true positives) and costs (false positives). ROC graphs are useful for visualizing, organizing, and selecting classifiers based on their comparative performance (Fawcett, 2006).

Evaluation of the COM

As mentioned in Section 2, the COM is a continuous classifier because it produces a continuous output, the health co-occurrence rate, which may be considered an estimate of the health-relatedness probability of queries. Each variant of the method has its own health co-occurrence rate with the distribution presented in the histograms of Figures 4, 5 and 6. In these histograms, only health co-occurrence rates between 0 and 1 are represented. In all three variants, we detected queries with health co-occurrence rates greater than 1: Google has 4,190, Yahoo! 769 while Yahoo!Google has 1,750 queries. Google has a co-occurrence average of 0.42, Yahoo! of 0.32 and Yahoo!Google of 0.38. The standard deviation is also greater in Google (0.28), followed by Yahoo!Google (0.23) and Yahoo! (0.22).

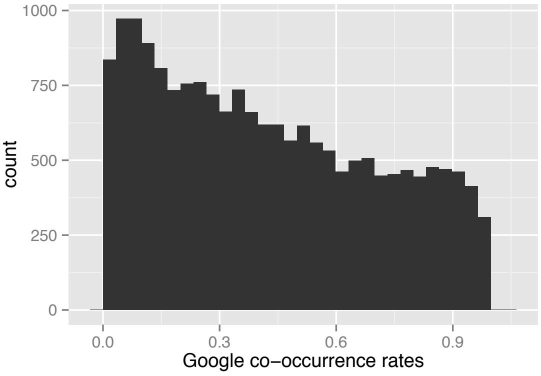

Figure 4 - Google health co-occurrence rate histogram

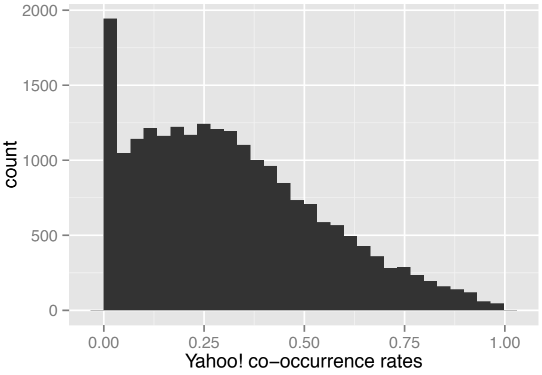

Figure 5 - Yahoo! health co-occurrence rate histogram

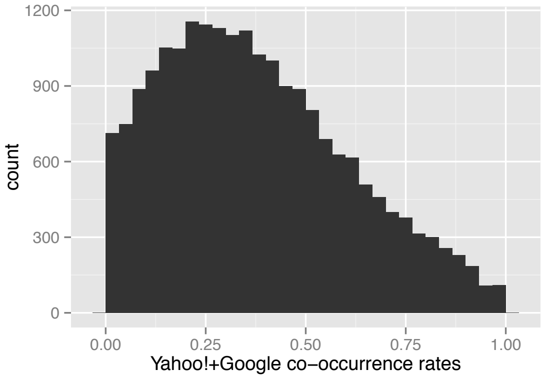

Figure 6 - Yahoo!Google health co-occurrence rate histogram

To predict each query health-relatedness, this continuous output was then compared with different thresholds (ranging from 0 to 1). Sensitivity, specificity, accuracy and the distance of each method to the optimal point (0,1) in the ROC space (ROCD) for the different thresholds in each method are presented in Table 2. Each column's maximum value is highlighted in bold with the exception of the last column, where the minimum value is the indicator of a best performance.

Table 2 - Sensitivity, specificity, accuracy and ROC distance for COM. Y – Yahoo!; G – Google; Y+G – Yahoo!Google

Thres-hold	SEN			SPE			ACC			ROCD		
	Y	G	Y+G	Y	G	Y+G	Y	G	Y+G	Y	G	Y+G
1	0.07	0.21	0.12	**0.97**	**0.82**	**0.93**	**0.92**	**0.78**	**0.88**	0.93	0.82	0.88
0.95	0.08	0.28	0.15	0.97	0.80	0.92	0.92	0.77	0.88	0.92	0.74	0.85
0.9	0.13	0.37	0.21	0.96	0.77	0.91	0.92	0.75	0.87	0.87	0.67	0.79
0.85	0.19	0.43	0.29	0.96	0.74	0.90	0.91	0.72	0.87	0.81	0.63	0.72
0.8	0.27	0.49	0.36	0.95	0.71	0.88	0.91	0.70	0.85	0.73	0.59	0.65
0.75	0.36	0.54	0.43	0.93	0.68	0.86	0.90	0.67	0.84	0.65	0.56	0.59
0.7	0.44	0.58	0.51	0.92	0.65	0.84	0.89	0.65	0.82	0.56	0.54	0.52
0.65	0.53	0.63	0.58	0.90	0.62	0.81	0.88	0.62	0.80	0.48	0.53	0.46
0.6	0.60	0.68	0.65	0.87	0.59	0.77	0.85	0.59	0.77	0.42	**0.52**	0.42
0.55	0.67	0.72	0.70	0.83	0.55	0.73	0.82	0.56	0.73	0.37	0.53	0.40
0.5	0.73	0.75	0.76	0.79	0.51	0.68	0.79	0.53	0.69	**0.34**	0.55	**0.40**

Threshold	**SEN**			**SPE**			**ACC**			**ROCD**		
	Y	**G**	**Y+G**	**Y**	**G**	**Y+G**	**Y**	**G**	**Y+G**	**Y**	**G**	**Y+G**
0.45	0.77	0.79	0.80	0.74	0.48	0.62	0.74	0.49	0.63	0.35	0.57	0.43
0.4	0.81	0.81	0.84	0.67	0.43	0.56	0.68	0.45	0.57	0.38	0.60	0.47
0.35	0.85	0.85	0.88	0.60	0.39	0.48	0.62	0.41	0.51	0.42	0.63	0.53
0.3	0.88	0.87	0.92	0.52	0.34	0.41	0.54	0.37	0.44	0.49	0.67	0.60
0.25	0.90	0.89	0.93	0.44	0.29	0.33	0.46	0.32	0.36	0.57	0.72	0.67
0.2	0.92	0.91	0.94	0.36	0.24	0.25	0.39	0.27	0.29	0.65	0.77	0.75
0.15	0.93	0.94	0.96	0.27	0.18	0.18	0.31	0.22	0.22	0.73	0.82	0.82
0.1	0.95	0.97	0.98	0.19	0.13	0.11	0.23	0.17	0.15	0.81	0.87	0.89
0.05	0.96	0.99	0.99	0.11	0.06	0.05	0.16	0.11	0.10	0.89	0.94	0.95
0	**1.00**	**1.00**	**1.00**	0.00	0.00	0.00	0.05	0.05	0.05	1.00	1.00	1.00

The ROC curves for each COM are presented in Figure 7. Each point in the curve corresponds to a threshold value, starting with 1 at the left-hand side of the graph.

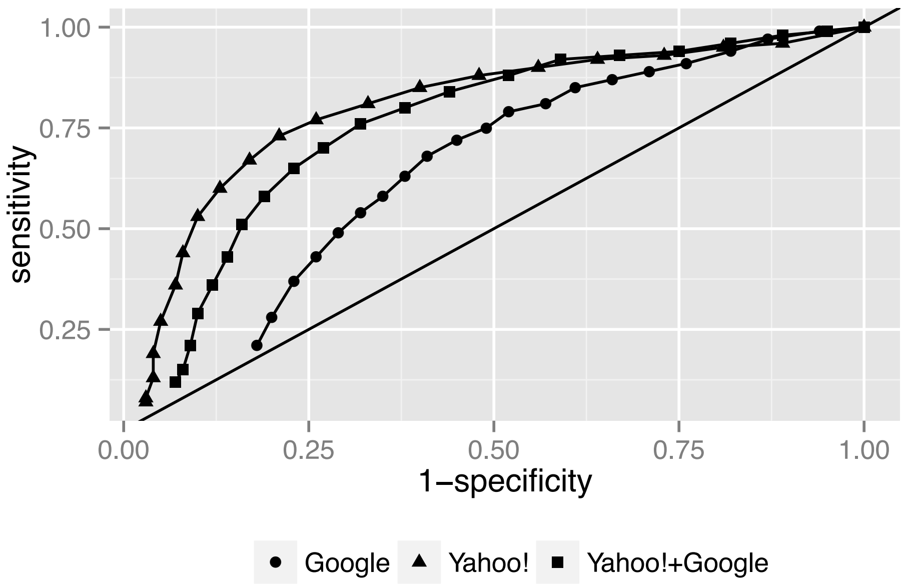

Figure 7 – COM ROC graph. The diagonal represents a random guess.

Evaluation of the CHV binary method

Table 3 presents, for each variant of the CHV binary method, the number of terms used in the classification method (Terms), the sensitivity, the specificity, the accuracy and the distance of each method to the optimal point (0,1) in the ROC space. Each column's maximum value is highlighted in bold with the exception of the last column, where the minimum value is the indicator of a best performance.

Table 3 - Number of terms, sensitivity, specificity, accuracy and ROC distance for the variants of the CHV binary method

Variant	**Terms**	**SEN**	**SPE**	**ACC**	**ROCD**
CHV 1	**158783**	**0.73**	0.35	0.37	0.71
CHV 2	1616	0.42	**0.85**	**0.83**	0.60

CHV 3	2897	0.51	0.80	0.79	0.53
CHV 4	4404	0.55	0.75	0.74	0.51
CHV 5	5622	0.57	0.73	0.72	**0.51**
CHV 6	20354	0.67	0.52	0.53	0.59
CHV 7	27657	0.43	0.73	0.72	0.63
CHV 8	58655	0.63	0.49	0.50	0.63
CHV 9	66398	0.65	0.48	0.49	0.63
CHV 10	5898	0.69	0.59	0.60	0.51
CHV 11	9872	0.71	0.52	0.53	0.56

To aid the comparison of the several variants, a ROC graph was drawn for each variant. This graph is presented in Figure 8.

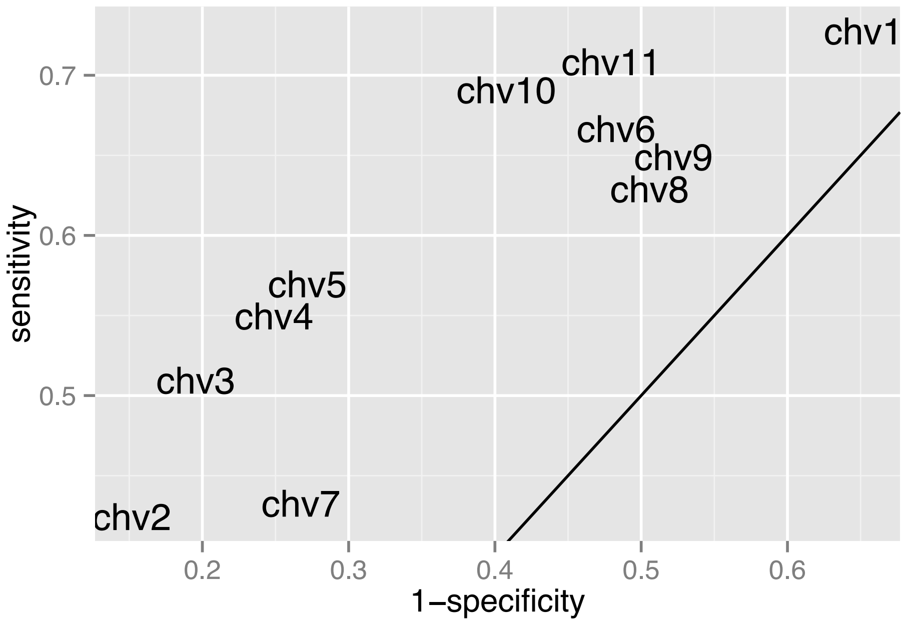

Figure 8 – CHV binary method ROC graph. The diagonal represents a random guess.

Evaluation of the CHV continuous method

As noted, we defined four CHV subsets that were used to do a preliminary evaluation of the CHV continuous method. The HEALTH subset included concept strings from UMLS categories containing concepts more likely to occur in consumer health queries, the CHVP contained the consumer preferred string for each CHV concept, the UMLSP had the UMLS preferred string for each CHV concept and the MEDP contained the MedlinePlus Health Topics source vocabulary concept strings.

A preliminary evaluation showed that the HEALTH subset produces the best results with respect to accuracy and distance to the ROC optimal point. However, the MEDP subset revealed a better specificity (86%-87%) due to a lower number of concept strings and its strong focus on consumers. In terms of sensitivity, M1Max using the UMLSP subset and M1Max using the CHV entire vocabulary had the best results with 68%. The UMLSP, despite having fewer strings than the CHV subset, has the same sensitivity probably because it contains almost all of the concept strings that led to the query classification. In general, almost all methods have sensitivity and accuracy values above 60%.

Table 4 shows the results of each variant of the method used in the classification of the sample collections in both Portuguese and English with the HEALTH Subset. As shown, the best variant is M2Max with a threshold of 0.17 using the English vocabulary. In Portuguese, the best variant is M1Max with a threshold of 0.5. We can therefore conclude that translation has impact on the results. The difference in sensitivity and accuracy is negligible. However, differences in the distance to the ROC optimal point and specificity are more expressive. We believe our results can be improved by removing unspecialized terms that, when used alone, are not health-related.

Table 4 - Best results with the HEALTH subset. T=threshold, L=language

Variant	T	L	SEN	SPE	ACC	ROCD
M1Max	0.2	EN	**0.76**	0.67	**0.73**	0.41
M1Avg	0.2		0.66	**0.80**	0.70	0.39
M1MaxBoost	0.2		0.71	0.71	0.71	0.41
M1AvgBoost	0.75		0.72	0.67	0.71	0.43
M2Max	0.17		0.68	0.79	0.71	**0.38**
M2Avg	0.1125		0.67	0.68	0.68	0.46
M2MaxBoost	0.35		0.71	0.71	0.71	0.41
M1Max	0.5	PT	0.65	0.69	**0.67**	**0.46**
M1Avg	0.2		0.65	0.68	0.66	0.47
M1MaxBoost	0.75		0.66	0.67	**0.67**	0.47
M1AvgBoost	0.2		0.67	0.65	0.66	0.48
M2Max	0.5		0.63	**0.70**	0.61	0.48
M2Avg	0.1		**0.68**	0.60	0.65	0.51
M2MaxBoost	0.75		0.66	0.67	0.66	0.47

6 - Discussion

Before analyzing the COM, we would like to mention the existence of health co-occurrence rates greater than 1. Theoretically, these values should not have existed as the default operator between terms in both search engines (Google and Yahoo!) is the logic "AND", implying that all terms in a query without operators should appear in the retrieved documents. In theory, adding terms should only result in a maintenance or decrease of the number of results. The number of queries in this situation is higher in Google than in Yahoo! (4,190 against 769). The query "go carts" is one example (with 3,230,000 results in Google) and the query "go carts health" (with 8,470,000 results in Google). This may be explained by the fact that the number of results returned by search engines is usually just an estimate. Google Help Center (2012) explains that not providing the exact count allows them to return search results faster. Yet, what is surprising still is the high number of these cases.

Figures 4, 5 and 6 show that the Yahoo!Google health co-occurrence rate is the closest to the Normal distribution, followed by the Yahoo! health co-occurrence rate. It is also possible to verify the existence of a surprising peak at the left side of the Yahoo! histogram. This peak shows that a large number of queries return 0 results.

Analyzing Table 2 data, it is possible to verify that, as expected, sensitivity is 1 at a threshold of 0. This happens because health co-occurrence rates are always bigger than 0 making all queries to be classified as health-related. Naturally, at this same threshold, specificity is 0 (as there are no queries classified as non-health related). Because of the generally high specificity values at the threshold of 1, accuracy is also maximized at this threshold. The analysis of the distance to the optimal point in the ROC space keeps the threshold of 0.5 as the best threshold of the Yahoo! method. Using Google, the best threshold value changes to 0.6 according to the same measure.

In Figure 7, the ROC graph clearly shows the dominance of Yahoo! over Google. Trend line for Yahoo! is always above Google's line. As well, it is possible to detect the closest points of each variant to the (0,1) point in the same figure. In ROC graphs, the point (0,1) represents a perfect classification, so better performances are closer to this point.

The idea of combining Yahoo! and Google estimates into a third method did not produced the improvements we expected with regard to the other two variants of this method. As shown in Figure 7 and Table 2, the Yahoo!Google variant has an intermediate performance, and is probably better than Google due to Yahoo! performance.

Google results in this sample of 20,000 queries are different from the results of Eysenbach and Kohler (2003). In their work, the threshold of 35% was considered an optimal trade-off between sensitivity (85.2%) and specificity (80.4%). The sample used in their study comprised 2,985 queries. Comparatively, our study had worse sensitivity values (68% or 72%), specificity values (59% or 55%) and different optimal threshold values (0.6 or 0.55). The larger sample used in our study makes us believe our results are a better portray of reality.

In Figure 8, it is possible to see all the variants of the CHV binary method to be better than a random guess, which is represented by a diagonal line, and those corresponding to the CHV binary method are located above it. Yet, no variant has reached the results initially expected. In fact, the best variants, as shown in Figure 8, are CHV2, CHV3, CHV4 and CHV5 (variants that use the list of terms of the 200, 400, 600 and 800 most frequent concepts) with their sensitivity not exceeding 57%. The specificity and accuracy is greater in CHV2, but sensitivity has a low value (42%) in this variant. CHV5 is the point closest to the (0,1) point.

We can also see that the number of health terms and sensitivity are not directly proportional. For example, CHV10 has fewer terms but a higher sensitivity and specificity than CHV6. This means that there are terms more related to the health context than others and that the performance of this method could be improved by a careful selection of terms. Generally, all the variants of the CHV binary method present a low sensitivity.

The results of the CHV continuous method in English show that this method outperforms all the variants of the CHV binary method and most of the COM variants. In fact, the best variant of the CHV continuous method has a ROC distance of 0.38 whereas the best variant of the CHV binary method has a ROC distance of 0.51 and the best COM variant achieved a distance of 0.34 with the Yahoo! search engine. The CHV continuous method has an extra advantage of allowing the association of queries with UMLS semantic types, which can improve the categorization of health queries.

The Portuguese results cannot be compared with the results of the other kind of methods, but allow us to conclude that although the translation has impact on the results, it can be a good strategy to apply CHV methods in non-English languages with further improvements in the translation process.

We should emphasize that the methods indicated as optimal may change if sensitivity is preferable to accuracy or vice-versa. For example, in a situation where we want to filter the number of queries to be categorized by a human assessor without the risk of eliminating a large number of health-related queries, it is preferable to have good sensitivity instead of specificity.

7 - Conclusions

In this paper, we evaluated three kinds of methods. Two of them were proposed by us and use terms from the CHV vocabulary. One produces a binary output and the other a continuous one. A third one was proposed by Eysenbach and Kohler (2003) and evaluates query relatedness to health through the co-occurrence rate of query terms with the word “health” in search engines’ results.

While Yahoo! performed better than Google in the COM, our results were worse than those reported by Eysenbach and Kohler (2003). In their work, at a threshold of 35%, sensitivity was 85.2% and specificity was 80.4%, while in our Yahoo! variant, we hit a threshold of 0.5 with 73% sensitivity and 79% specificity. We believe our results depict reality more accurately as our sample of queries is an order of magnitude larger: 20,000 against 2,985 queries.

None of the binary methods that used subsets of terms of health vocabularies behaved as well as the Yahoo! variant. Yet, some variants of the CHV binary method behaved better than the Google variant (CHV3, CHV4 and CHV5 had better or similar performance than the Google variant).

In summary, the variants of the CHV continuous method outperformed most of the other methods. Equally important is the fact that this method allows the association of queries to the UMLS semantic tree and their classification into categories like *Disease or Syndrome* or *Anatomical Structure*. The output of our method can be useful to search engines, for example, search engines can apply our method to provide contextualized query suggestions or even information about the health subject being sought.

Finally, the evaluation of this set of methods in a language other than the vocabulary language showed that the influence of the translation process in the proposed method may be noticeable, but it does not generally compromise its overall effectiveness.

References

Beitzel, S., Jensen, E., Frieder, O., Lewis, D., Chowdhury, A., & Kolcz, A. (2005). *Improving Automatic Query Classification via Semi-Supervised Learning.* Paper presented at the Proceedings of the Fifth IEEE International Conference on Data Mining.

Chitu, A. (2007). Google Finds Less Search Results. Retrieved 12 March, 2014, from http://googlesystem.blogspot.pt/2007/12/google-finds-less-search-results.html

Eysenbach, G., & Kohler. (2003). *What is the prevalence of health-related searches on the World Wide Web? Qualitative and quantitative analysis of search engine queries on the internet*. Paper presented at the AMIA Symposium. http://view.ncbi.nlm.nih.gov/pubmed/14728167

Fawcett, T. (2006). An introduction to ROC analysis. *ROC Analysis in Pattern Recognition, 27*(8), 861-874. doi: 10.1016/j.patrec.2005.10.010

Fox, S. (2011). Health Topics.

Google. (2012). Google search result count. Retrieved 2014-03-12, from http://support.google.com/webmasters/bin/answer.py?hl=en&answer=70920

Kleinsorge, R., & Willis, J. (2007). Unified Medical Language System (UMLS) Basics.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). Introduction to Latent Semantic Analysis. *Discourse Processes*(25), 259-284. doi: citeulike-article-id:13118621

Lopes, C., Dias, D., & Ribeiro, C. (2013). Using Domain-Specific Term Frequencies to Identify and Classify Health Queries. In Á. Rocha, A. Correia, T. Wilson & K. Stroetmann (Eds.), *Advances in Information Systems and Technologies* (Vol. 206, pp. 221-226): Springer Berlin Heidelberg.

McCray, A. T., Loane, R. F., Browne, A. C., & Bangalore, A. K. (1999). *Terminology issues in user access to Web-based medical information*. Paper presented at the AMIA Symposium. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2232498/

Murata, T. (2007). Detection of Breaking News from Online Web Search Queries. *New Generation Computing, 26*(1), 63-73. doi: 10.1007/s00354-007-0035-3

Nlm. (2012). 2012AA Consumer Health Vocabulary Source Information.

Spink, A., Wolfram, D., Jansen, M., & Saracevic, T. (2001). Searching the Web: the public and their queries. *J. Am. Soc. Inf. Sci. Technol., 52*(3), 226-234. doi: citeulike-article-id:1189994
doi: 10.1002/1097-4571(2000)9999:9999%3C::aid-asi1591%3E3.3.co;2-i

Spink, A., Yang, Y., Jansen, J., Nykanen, P., Lorence, D. P., Ozmutlu, S., & Ozmutlu, H. C. (2004). A study of medical and health queries to web search engines. *Health Information & Libraries Journal, 21*(1), 44-51. doi: 10.1111/j.1471-1842.2004.00481.x

Zeng, Q., Crowell, J., Plovnick, R., Kim, E., Ngo, L., & Dibble, E. (2006). Assisting consumer health information retrieval with query recommendations. *Journal of the American Medical Informatics Association : JAMIA, 13*(1), 80-90. doi: 10.1197/jamia.m1820