Using Domain-specific Term Frequencies to Identify and Classify Health Queries

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

[1] DEI, Faculdade de Engenharia, Universidade do Porto
[2] INESC TEC
Rua Dr. Roberto Frias s/n, 4200-465, Portugal
{ctl,ei06025,mcr}@fe.up.pt

Abstract. In this paper we propose a multilingual method to identify health-related queries and classify them into health categories. Our method uses a consumer health vocabulary and the Unified Medical Language System semantic structure to compute the association degree of a query to medical concepts and categories. This method can be applied in different languages with translated versions of the health vocabulary. To evaluate its efficacy and applicability in two languages we used two manually classified sets of queries, each on a different language. Results are better for the English sample where a distance of 0.38 to the ROC optimal point (0,1) was obtained. This shows some influence of the translation in the method's performance.

Keywords: Health Information Retrieval, Health Queries, Medical Vocabularies, Web Information Retrieval.

1 Introduction

The Web is now a major source of information worldwide and the use of search engines to find health information is a common practice. In 2006, 80% of Internet users in the United States used the Web to search for health information [3]. According to Eysenbach and Köhler, over 12 million health queries are made per day in *Google* [2]. The classification of queries is frequently used to distinguish them according to the topic. This classification can be manual, may involve the comparison of a query with databases of queries or require machine learning processes. Another possibility is to use controlled vocabularies or thesaurus of terms, in areas where the quality of these structures can be trusted. Since most health queries contain terms that can be mapped to health vocabularies [5], we propose a method to detect consumer health queries that takes advantage of existing high-quality health vocabularies, can be applied in different languages and can classify queries into health categories like diseases.

2 Related work

Two previous works report methods to identify health queries. Eysenbach and Köhler [2] 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. In another work, Lopes [4] compares the Eysenbach and Köhler's method with a method that uses health vocabularies to identify health queries. In this last method, the author considers that the presence of a health vocabulary's term in a query is sufficient to classify the query as being health-related. In this work, several variants of both methods are compared.

Like the work of Lopes [4], our work will use health vocabularies but in a different way. While the previously described method is discrete, simply indicating if a query is or not a health query, our method will produce a score, indicating the degree to which a query is related to the health domain. Moreover, our method can be used to classify health queries into categories like *Disease or Syndrom* or *Anatomical Structure*.

3 Proposed method for query classification

The proposed method takes advantage of the UMLS predefined structures and the Consumer Health Vocabulary (CHV). It includes the creation of indexes to help the comparison between query terms and the health vocabulary and the calculation of the final score. Besides classifying each query as being health-related or not, we also associate it with the UMLS specialized health categories.

3.1 Health Semantic Structures

We have chosen the Consumer Health Vocabulary[3] (CHV), developed under an open source and collaborative initiative that is linked to the Unified Medical Language System (UMLS) and its many sources. The CHV has 42,977 health concepts and 158,508 concept strings in English. The UMLS is one of the most consistent and robust health semantic structures including about 1 million biomedical concepts from 100 different sources and a large semantic structure.

3.2 Vocabulary Translation

One of the main disadvantages of a method based on vocabularies is its dependence on the language and country in which it was created. Our hypothesis here is that we can apply our method by previously translating the CHV without much penalty on the results. We expect the translation process to have some influence on the classification results but also hope to minimize it using a good translation process. To evaluate the efficacy of our method in a language other than English we used the *Google Translator API*. We manually evaluated 1% of the total number of translated strings and concluded that 84.2% of the translations were good, which is very satisfactory.

[3] http://www.consumerhealthvocab.org

3.3 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 decided to obtain different subsets of the CHV vocabulary instead of using only 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 category concept strings (MEDP). MedlinePlus is a website for health consumers and the UMLS category with this name contains the concepts explored in this site.

3.4 Auxiliary Structures

For each subset, we created an inverted index containing the unique terms mapped to a list of unique identifiers for each concept string in the subset and their association degree with each 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 in 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.

3.5 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 these lists to calculate the final score for the query. As showed in Figure 1, 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 $w_c^q \times \frac{cf_{c,q}}{|q|}$, where $cf_{c,q}$ is the frequency of c in query q and $|q|$ the number of unique terms in q.

3.6 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, presented in Table 1. Here the $Query$ is the query list obtained after the previous combination in each method, $tf_{h,q}$ is the number

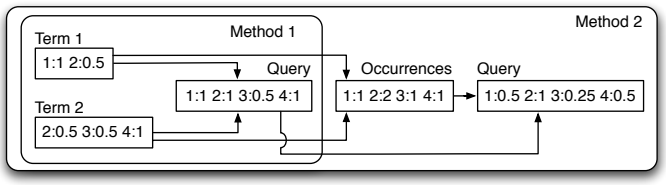

Fig. 1. Joining posting lists in Methods 1 and 2.

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	$\max(Query) \times (tf_{h,q} \div \|q\|)$	No
M1MaxBoost		Yes
M1Avg	$avg(top_1^5(Query)) \times (tf_{h,q} \div \|q\|)$	No
M1AvgBoost		Yes
M2Max	$\max(Query)$	No
M2MaxBoost		Yes
M2Avg	$avg(Query)$	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. An example of this case is the query "tooth piercing" as "tooth" is a full concept in the CHV subset as an anatomical part and the term "piercing" doesn't appear in it. Without the final product this query would have a score of 1 and it scores 0.5 with it. 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 Information Retrieval and is equal to the number of strings in the CHV in which the term appears.

3.7 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. In this sense 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.

4 Findings and discussion

To evaluate the methods using the English (EN) CHV concept strings we have used a dataset created by Beitzel and Lewis who had queries classified into 20 topical categories by a team of approximately ten human assessors. We included 1,647 queries, part classified as health queries and part classified into other categories [1]. In Portuguese (PT) we have used a collection of 1,522 queries manually classified by medical students. For each method we calculated the true positive rate (TPR), false positive rate (FPR), accuracy (ACC) and the distance (ROCD) to the optimal point in the ROC Space (0,1).

With all the CHV subsets, initial tests 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 FPR (13%-14%) due to a lower number of concept strings and its focus on consumers. In terms of TPR, 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 TPR probably because it contains almost all of the concept strings that led to query classification. In general, almost all methods have TPR and ACC values above 60%.

Table 2 shows the results of each method used in the classification of the sample collections in both languages with the HEALTH Subset. As shown, the best method is M2Max with a threshold of 0.17 using the English vocabulary. In Portuguese the best method is M1Max with a threshold of 0.5. We can therefore conclude that translation has impact on the results. The difference in TPR and ACC is negligible. However, differences in ROCD and FPR are more expressive. We believe our results can be improved by removing unspecialized terms that, alone, are not health-related.

Comparing our results in the English language with the results obtained by Lopes [4], we notice that our best method has a ROCD of 0.38, a little worse than Lopes's best result. Her best result was obtained applying the Eysenbach and Köhler method using the Yahoo! search engine and had a ROCD of 0.34. However, our method has a smaller ROCD than all the other variants of the Eysenbach and Köhler method and all the methods that use health vocabularies. Moreover it has the advantage of being able to associate the queries with the UMLS specialized health categories.

Table 2. Best results in the HEALTH subset. T=threshold, L=language.

M	T	L	TPR	FPR	ACC	ROCD
M1Max	0.2	EN	**0.76**	0.33	**0.73**	0.41
M1Avg	0.2	EN	0.66	**0.2**	0.7	0.39
M1MaxBoost	0.2	EN	0.71	0.29	0.71	0.41
M1AvgBoost	0.75	EN	0.72	0.33	0.71	0.43
M2Max	0.17	EN	0.68	0.21	0.71	**0.38**
M2Avg	0.1125	EN	0.67	0.32	0.68	0.46
M2MaxBoost	0.35	EN	0.71	0.29	0.71	0.41
M1Max	0.5	PT	0.65	0.31	**0.67**	**0.46**
M1Avg	0.2	PT	0.65	0.32	0.66	0.47
M1MaxBoost	0.75	PT	0.66	0.33	**0.67**	0.47
M1AvgBoost	0.2	PT	0.67	0.35	0.66	0.48
M2Max	0.5	PT	0.63	**0.30**	0.61	0.48
M2Avg	0.1	PT	**0.68**	0.40	0.65	0.51
M2MaxBoost	0.75	PT	0.66	0.33	0.66	0.47

5 Conclusions

This work proposes a new method to identify and classify health-related queries that explores the UMLS predefined structures and can be applied in different languages. The influence of the translation process in the proposed method is noticeable but does not compromise its overall effectiveness. Moreover, and not less important, our approach allows the association of queries to the UMLS semantic tree and their classification into categories like *Disease or Syndrom* or *Anatomical Structure*. The output of our method can be useful to search engines that can, for example, use it to provide contextualized query suggestions or even information about the health subject searched for. In the future, we would like to test these methods with an inverted index created with multiple data from different vocabularies and to combine them with machine learning techniques.

References

1. Beitzel, S.M., Jensen, E.C., Frieder, O., Lewis, D.D., Chowdhury, A., Kolcz, A.: Improving Automatic Query Classification via Semi-Supervised Learning. In: Fifth IEEE International Conference on Data Mining (2005)
2. Eysenbach, G., Köhler, C.h.: What is the prevalence of health-related searches on the World Wide Web? Qualitative and quantitative analysis of search engine queries on the internet. AMIA Symposium (2003)
3. Fox, S.: Online Health Search 2006. Tech. rep., Pew Internet & American Life Project (2006)
4. Lopes, C.T.: Evaluation and comparison of automatic methods to identify health queries. In: Doctoral Symposium on Informatics Engineering 2008 (2008)
5. Zeng, Q.T., Crowell, J., Plovnick, R.M., Kim, E., Ngo, L., Dibble, E.: Assisting consumer health information retrieval with query recommendations. JAMIA (2006)