# Exploiting Health Related Features to Infer User Expertise in the Medical Domain

João R. M. Palotti
Vienna University of Technology
palotti@ifs.tuwien.ac.at

Allan Hanbury
Vienna University of Technology
hanbury@ifs.tuwien.ac.at

Henning Müller
University of Applied Sciences and Arts Western Switzerland
henning.mueller@hevs.ch

## ABSTRACT

The internet is an important source of health knowledge for everyone, from laypeople to medical professionals. It is known that these two groups have distinct needs and distinguishing them can significantly improve their satisfaction. In this work, the logs of two web search engines are augmented with annotations provided by the US National Library of Medicine (NLM) tool Metamap and various features are created and applied in a user expertise classification task. We focus on generating features relevant to what the users search for instead of how they search. The results showed that a classifier using the health related features proposed can boost the classification accuracy by more than 14%, compared to the same classifier using only basic user behaviour features.

## Categories and Subject Descriptors

H.3 [**Information Systems**]: Information Storage and Retrieval; J.3 [**Life and Medical Sciences**]: Medical Information Systems

## General Terms

Algorithms, Experimentation

## Keywords

Query log analysis, health search, expertise prediction

## 1. INTRODUCTION

Health is one of the most important and searched topics on the internet. According to a recent survey, one in three American adults went online to diagnose some medical condition they or someone else might have [6]. Not only patients use the internet, physicians are active internet users as well. PubMed, which indexes the biomedical literature, reports more than one hundred million users of which two-thirds are medical professionals [9].



Distinguishing between experts and laypeople can improve significantly the user's interactions with the search engine. Currently, users may get different results for their queries if they are in different locations, but few (if any) change is seen if they are close to each other, but have different levels of expertise. In the health domain, this often leads to experts struggling to get advanced results on a disease, while laypeople cannot understand the same results due to lack of technical vocabulary, for example. White et al. [11] suggest that search engines could provide non-experts with definitions for expert terms, or even teach laypeople to identify reliable sites that fit their reading skills, potentially displaying more complex information as the novice gains knowledge.

The detection of experts has attracted the attention of researchers, but mainly domain independent features were studied [11, 14, 4]. White et al. [11] is an important related work, as the authors also deal with analysis of query logs. They assume that searches leading to PubMed were made by medical experts and searches leading to ACM Digital library (ACM-DL) were made by computer science experts. In the medical domain this is a weak premise for two reasons: (1) it is estimated that one-third of PubMed users are laypeople [9], (2) PubMed is more important for medical researchers than practitioners [8]. Tracing a parallel between medicine and computer science, a General Practitioner would be like a software developer that does not necessarily need to consult the ACM-DL (the correspondent for PubMed) to perform his/her work. One could manually expand the list of expert sites to include, for example, StackOverflow or an API website for experts in CS and treatment guidelines or drug information sites for medicine, but it would be a laborious task and instable over time. Hence, defining the ground truth itself is already a complex problem in the expertise detection task.

To cope with this issue we use the logs of two search engines made for distinct audiences: (1) TRIP[1] for medical professionals searching for clinical evidence, and (2) HON[2] for patients searching for trustworthy material. Both search engines present a user interface similar to standard web search engines, taking free text queries in a single text box. Our assumption does not require any complex filtering of users, as most users query medical content in these websites.

One concern that arises when using two different sources of logs is that we could learn how to differentiate between the two search engines, instead of learning how to infer the correct user expertise. To overcome this drawback, we are

[1] http://www.tripdatabase.com/
[2] http://www.hon.ch/HONsearch/Patients/index.html

focusing on *what* the users search for (e.g., analysing the keywords used to find out the topics searched), rather than *how* (e.g., number of words used or number of queries per session). Therefore, we do not make use of user sessions in this work, rather we focus on creating and evaluating features for expertise prediction in the health domain, based solely on the keywords used. As an outcome, we built a classification model capable of inferring user medical expertise that can be easily integrated into any search engine. The results show that a Random Forest classifier using the medical features proposed can boost the classification accuracy by more than 14%, compared to the same classifier using only user behavior related features.

## 2. RELATED WORK

The difference between the use of search engines by experts and laypeople has been long studied in the literature. In the 1990's, for example, Hsieh-Yee [7] reported that experienced library science students could use more thesauri, synonymous terms, combinations of search terms and expend less time monitoring their searches than novices. Later, Bhavnani [2] studied search expertise in the medical and shopping domain. He reported that experts in a topic can easily solve the task given even without using a search engine, because they already knew which website was more propitious to fill their needs. Duggan and Payne [5] explored the domains of music and football to evaluate how the user knowledge of a topic can influence the probability of a user answering factual questions, finding that experts give up a barren line of inquiry faster than non-experts.

Similarly to our work, White et al. [11] built a classifier to predict user expertise in 4 domains (medicine, finance, law, and computer science) based on the query logs of a general search engine. Our work differs from White's in many aspects. We employed different query logs used by distinct target audiences, while they filtered results considering experts any user who had accessed one of a few hand-picked websites. Also, we focus on medical queries only, giving us the opportunity to explore and evaluate more domain specific features, rather than traditional user behaviour features.

More recent studies in user expertise prediction include Zhang et al. [14] and Cole et al. [4], both user studies ($|n| \leq 40$) using TREC Genomics as the dataset. The former employed a regression model to match user self-rated expertise, and high level user behavior features such as the mean time analysing a document and the number of words per query. Similarly, but using only eye movement patterns as features, the latter employed a linear model and random forests to infer the user expertise level.

## 3. DATA COLLECTION

Two query logs from search engines taking free text queries were used: one focused on laypeople queries and the other consists of queries from medical professionals. Some general statistics about the data used are described in Table 1.

The query logs assumed to consist almost completely of queries submitted by laypeople were obtained from the Health of the Net Foundation website (HON). This non-governmental organization is responsible for the HONcode, a certification of quality given to websites fulfilling a pre-defined list of criteria. They provide a search engine to facilitate the access to the certified sites. 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[3].

As the professional dataset, we are using the logs from the Turning Research Into Practice (TRIP) database. It is a search engine for medical evidence 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 [10].

Table 1: General statistics for the data used in this work.

| Dataset | HON | TRIP |
|---|---|---|
| **Users** | 92,111 | 279,506 |
| **Total Queries** | 343,007 | 1,853,233 |
| **Unique Queries** | 228,121 | 541,979 |
| **Initial Date** | December 2011 | January 2011 |
| **Final Date** | August 2013 | August 2012 |

Only two basic pieces of information were extracted for each query: (1) the anonymous user identification, and (2) the keywords used. This information is the common intersection of the two query logs used, and potentially present in any other query log of a search engine. Then, we enrich this simple information using Metamap [1], a well-known tool from the US National Library of Medicine (NLM). We used the default processing options of Metamap to generate various features based only on the keywords of each query. We augment the datasets with: (1) the concepts found in each query, (2) the sources of vocabularies, (3) the Medical Subject Headers (MeSH) identifiers, (4) the medical semantic types, and (5) the part-of-speech tagging. In the next section, we describe the 28 features generated in this work, many of them used for the first time in this task.

## 4. FEATURE DESCRIPTION

We divided the features created into 4 groups to better organize the features and analyse the contribution of each group. All the groups are described below and a summary is presented in Table 2.

### 4.1 Semantic Features

Based on the work of Cartright et al. [3], three semantic classes were created throughout the semantic types given by Metamap for each query: (1) **symptom**, (2) **cause** (disease), (3) **remedy**. For example, a search containing "bipolar disorder" is annotated by Metamap with the type *Mental or Behavioral Dysfunction*, which is attributed to the semantic meaning **cause**. The complete list of semantic types Metamap can produce is available online[4], while the mapping from Metamap's semantic types to Cartright's symptom-cause-remedy types is the following (with examples in parenthesis):

- **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)

[3]The Latin 1 covers the majority of European languages, however it excludes the majority of Asian languages

[4]`http://metamap.nlm.nih.gov/Docs/SemanticTypes_2013AA.txt`

- **Remedy:** clnd (gatorade; cough syrup), antb (antibiotic; penicillin), aapp(vectibix; degarilex), phsu (tylenol; mietamizol), imft (vaccine; acc antibody), vita (vitamin B12; quercetin)

A query with no symptom, cause or remedy is attributed to the type **other** (Avg. Other Types Per Query).

## 4.2 UMLS Features

The Unified Medical Language System (UMLS) Metathesaurus is a multi-purpose, multilingual vocabulary database, containing information about biomedical and health related concepts. It is updated quarterly with new vocabularies and currently contains 169 different sources[5]. Altogether, UMLS comprises more than 1 million biomedical concepts.

Metamap provides an easy way to access all the sources and different concepts to which a term may belong. Using this information, we model features such as the average number of sources per query and the average number of concepts used by a user. It is also easy to map each UMLS concept to one or more concepts in the Medical Subject Headings (MeSH[6]) hierarchy. MeSH was already used assuming that difficult concepts are lower in the hierarchy [12].

## 4.3 Consumer Health Vocabulary Features

The vocabulary gap between laypeople and professionals is a substantial barrier to health information access for laypeople. The Consumer Health Vocabulary (CHV) was created to cope with this issue [13]. The CHV dataset (version 20110204) links part of the UMLS concepts, such as "myocardial infarction", to everyday expressions, "heart attack". Moreover, for many terms in the UMLS Metathesaurus, a set of three difficulty scores is available, related to the frequency or the context in which the term is used. We used only the *Combo Score*, which combines the other two scores. For any word without a *Combo Score*, we used the mean *Combo Score* of the complete CHV dataset, 0.29 (the data ranges from 0.0 - very difficult - to 1.0 - very easy).

For each query in our datasets, we compute five values: (1) the number of terms found in the CHV dataset; if the query contained an (2) expert term, a (3) layperson term or a (4) misspelled term; as well as (5) the average *Combo Score* of all terms identified. Therefore, the query "heart attack" contains only one term in the CHV dataset, which is a laypeople term, and its difficulty score is 0.8.

## 4.4 Part-of-Speech Tagging Features

We employed the module MedPost/SKR of Metamap to annotate each word in a query with one of the following lexicon tags: noun, verb, auxiliary verb, adjective, conjunction, adverb, determiner, preposition, modal verb, pronoun, punctuations and shapes (numbers).

# 5. EXPERIMENTS

The classification problem presented here seeks to infer the user expertise based on his/her queries by calculating the features showed in Table 2 for each user. As shown in Table 1, there are 92,111 regular users and 279,506 medical users in the dataset, resulting in a baseline accuracy of 75.21% for a classifier that assigns all the users to the most frequent class (MFC), the medical professionals. We are aware that this dataset distribution does not necessarily represents the real world distribution. Instead of sampling our dataset, we decide to make use of all users available and report metrics that take into account class imbalance.

Table 2: Each one of the 4 groups and the 28 features used

| | |
|---|---|
| **Semantic Features** | |
| Avg. Symptoms Per Query | Avg. Causes Per Query |
| Avg. Remedies Per Query | Avg. Other Types Per Query |
| **Unified Medical Language System (UMLS) Features** | |
| Avg. Queries Using Sources | Avg. Sources Per Query |
| Avg. Queries Using Concepts | Avg. Concepts Per Query |
| Avg. Queries Using MeSH | Avg. MeSH Per Query |
| Avg. MeSH Depth Per Query | |
| **Consumer Health Vocabulary (CHV) Features** | |
| Avg. CHV Terms Per Query | Avg. Layperson Terms Per Query |
| Avg. Expert Terms Per Query | Avg. Misspelled Terms Per Query |
| Avg. Combo Score Per Query | |
| **Part-of-Speech Tagging (POS) Features** | |
| % of Nouns | % of Verbs |
| % of Auxiliary Verbs | % of Adjectives |
| % of Conjunctions | % of Adverbs |
| % of Determiners | % of Prepositions |
| % of Pronouns | % of Shapes |
| % of Punctuations | % of Modal Verbs |

Naive Bayes, Support Vector Machines, Logistic Regression and Random Forest (RF) from the Python package scikit-learn[7] were used with their respective best hyperparameters selected using grid search. We show here only the results for the RF, the classifier which had the best performance. The performance of each classifier was measured by precision, recall, $F_1$[8] and accuracy scores, as those are well known and widely used metrics. We also report the Mean True Positive Rate ($\mu_{TPR}$) defined as: $\mu_{TPR} = 100 \times \frac{TPR_{c_1}+TPR_{c_2}}{2}$, where $TPR_{c_x}$ is the true positive rate for class $c_x$. We investigate $\mu_{TPR}$ because the distribution of the classes is unbalanced and this metric forces the MFC classifier to score exactly 50 out of 100, making easier and more understandable the comparison of different methods.

We performed a ten-fold cross-validation experiment across ten runs. We analyse the results of our model and compare them to two baselines: (1) a classifier that assigns all examples to the positive class (accuracy and $\mu_{TPR}$ are calculated using the most frequent class as the positive class), (2) a Random Forest classifier using two basic user behaviour related features - avg. words per query and avg. characters per query. These two features were chosen because they have shown to be good predictors on the literature [11, 14].

Table 3 summarizes the results. We use the term *medical features* to refer to the features described in Section 4 and *basic features* to the two user behaviour related features mentioned above. Rows 1 and 2 report the two baselines proposed, while rows 3 and 4 are the results of using only the medical features, and using all features (medical + ba-

[5] `http://www.nlm.nih.gov/research/umls/sourcereleasedocs/index.html`

[6] MeSH is a controlled hierarchical vocabulary used by NLM for indexing journal articles in the life sciences field. The whole hierarchy contains more than 25,000 of subject headings, with the most recent version containing 16 top categories such as "Anatomy" and "Diseases"

[7] `scikit-learn.org`

[8] $F_1$ is defined as: $F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$

Table 3: All the models were compared to the model in the previous row yielding a $p < 0.01$ in a two tailed t-test

| # | Classifier | Pos. Class | Acc. | Prec. | Rec. | $F_1$ | $\mu_{TPR}$ |
|---|---|---|---|---|---|---|---|
| 1 | **Weak Baseline** | Layp. | 75.21 | 24.79 | 100.00 | 39.73 | 50.00 |
| | Positive Class | Exp. | | 75.21 | 100.00 | 85.85 | |
| 2 | **Random Forest** | Layp. | 76.00 | 53.91 | 21.95 | 31.20 | 57.88 |
| | Basic Features | Exp. | | 76.00 | 93.82 | 85.47 | |
| 3 | **Random Forest** | Layp. | 87.23 | 80.12 | 64.50 | 71.46 | 79.61 |
| | Medical Features | Exp. | | 89.01 | 94.72 | 91.78 | |
| 4 | **Random Forest** | Layp. | 87.78 | 82.05 | 64.92 | 72.49 | 80.12 |
| | All Features | Exp. | | 89.18 | 95.32 | 92.15 | |

sic). Row 2 shows that there is only little improvement in accuracy or $F_1$ when using a RF and basic features, mainly because of the poor recall for detecting users belonging to the laypeople class. In row 3 we see that when using the medical features, the RF classifier got statistically significant gains over both baselines, reaching an improvement of 14% in accuracy.

The RF classifier also allows us to compute the Gini importance score for each feature. This value (from 0 to 100) 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. Table 4 shows the top ten features according to the Gini importance score when all the features are used. Although the RF using only the two basic features did not perform well, these features are among the top five. Unfortunately, the high variance of this method does not allow us to say that one single feature is statistically better than the others. Due to space constraints a more detailed feature study is left as future work.

Table 4: Top 10 features according to the Gini importance score generated by the Random Forest classifier

| Rank | Feature Name | Gini | Group |
|---|---|---|---|
| 1 | Avg. CHV Terms | 11.43 ± 7.04 | CHV |
| 2 | Avg. Combo Score | 11.19 ± 4.44 | CHV |
| 3 | Avg. Chars Per Query | 9.49 ± 0.55 | Basic |
| 4 | Avg. Sources Per Query | 7.79 ± 2.86 | UMLS |
| 5 | Avg. Words Per Query | 5.74 ± 0.64 | Basic |
| 6 | Avg. Concepts Per Query | 5.40 ± 3.03 | UMLS |
| 7 | Avg. MeSH Depth | 5.17 ± 1.24 | UMLS |
| 8 | Avg. Expert Terms Per Query | 4.93 ± 4.93 | CHV |
| 9 | Percentage of Nouns | 4.28 ± 0.88 | POS |
| 10 | Avg. MeSH Per Query | 3.94 ± 1.15 | UMLS |

A direct comparison with other works in the literature such as White et al. [11], Zhang et al. [14] and Cole et al. [4] would not be fair, because these other works use a different range of features and datasets. Particularly, many of the features are related to the result page: ranking of clicked results, domains of results, saved documents, among others. In contrast to the metrics used in this work, these are pieces of information more difficult to obtain and not always available in search logs.

# 6. CONCLUSIONS AND FUTURE WORK

In this work we have developed and evaluated features to be used in the medical domain to classify users according to their expertise. We concentrated on pieces of information easily obtained by any search engine: the keywords for each query. Many of the features have never been used in the literature before, and altogether it was possible to outperform the two baselines, reaching an accuracy of 87%. As future work we expect to evaluate not only keywords, but the user's sessions. We also plan to apply the classification model proposed here to ranking documents differently, providing query suggestions and supporting different levels of readability based on the user expertise.

**ACKNOWLEDGMENTS.**

This research was funded by the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement n°257528 (KHRESMOI) and partly funded by the Austrian Science Fund (FWF) project number I1094-N23 (MUCKE).

# 7. REFERENCES

[1] A. R. Aronson. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. In *Proc AMIA Symp*, 2001.

[2] S. K. Bhavnani. Domain-specific search strategies for the effective retrieval of healthcare and shopping information. Proc. SIGCHI, 2002.

[3] M.-A. Cartright, R. W. White, and E. Horvitz. Intentions and attention in exploratory health search. In *Proc of SIGIR*, 2011.

[4] M. J. Cole, J. Gwizdka, C. Liu, N. J. Belkin, and X. Zhang. Inferring user knowledge level from eye movement patterns. *Inf. Process. Manage.*, 2013.

[5] G. B. Duggan and S. J. Payne. Knowledge in the head and on the web: using topic expertise to aid search. Proc. SIGCHI, 2008.

[6] S. Fox and M. Duggan. Health online 2013. Technical report, The Pew Internet & American Life Project, January 2013.

[7] I. Hsieh-Yee. Effects of search experience and subject knowledge on the search tactics of novice and experienced searchers. *JASIS*, 1993.

[8] M. Kritz, M. Gschwandtner, V. Stefanov, A. Hanbury, and M. Samwald. Utilization and perceived problems of online medical resources and search tools among different groups of european physicians. *J Med Internet Res*, Jun 2013.

[9] E.-M. Lacroix and R. Mehnert. The US National Library of Medicine in the 21st century: expanding collections, nontraditional formats, new audiences. *HILJ*, 2002.

[10] 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*, 2007.

[11] R. W. White, S. T. Dumais, and J. Teevan. Characterizing the influence of domain expertise on web search behavior. In *Proc. of WSDM*, 2009.

[12] X. Yan, R. Y. Lau, D. Song, X. Li, and J. Ma. Toward a semantic granularity model for domain-specific information retrieval. *Trans. Inf. Syst.*, 2011.

[13] Q. T. Zeng and T. Tse. Exploring and developing consumer health vocabularies. *JAMIA*, 2006.

[14] X. Zhang, M. Cole, and N. Belkin. Predicting users' domain knowledge from search behaviors. Proc. SIGIR, 2011.