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

Carla Teixeira Lopes, Department of Informatics and Engineering, University of Porto, Porto, Portugal
Cristina Ribeiro, Department of Informatics and Engineering, University of Porto, Porto, Portugal and INESC Technology and Science, Institute for Systems and Computer Engineering of Porto, Porto, Portugal

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, Health Vocabularies, Information Filtering, Information Retrieval

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.

DOI: 10.4018/ijhisi.2014070104
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 article 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
Related Content

Multi-sensing Monitoring and Knowledge-driven Analysis for Dementia Assessment
[www.igi-global.com/article/multi-sensing-monitoring-and-knowledge-driven-analysis-for-dementia-assessment/134012?camid=4v1a](www.igi-global.com/article/multi-sensing-monitoring-and-knowledge-driven-analysis-for-dementia-assessment/134012?camid=4v1a)

Ethnographic Discovery of Adverse Events in Patient Online Discussions: Customer Relationship Management
[www.igi-global.com/article/ethnographic-discovery-adverse-events-patient/2233?camid=4v1a](www.igi-global.com/article/ethnographic-discovery-adverse-events-patient/2233?camid=4v1a)
Opportunities and Barriers of Sexual Health and Condom Use among Tea Plantation Workers
[www.igi-global.com/article/opportunities-barriers-sexual-health-condom/54020?camid=4v1a](www.igi-global.com/article/opportunities-barriers-sexual-health-condom/54020?camid=4v1a)

Expanding Role of Telephone Systems in Healthcare: Developments and Opportunities
Jing Shi, Ergin Erdem and Heping Liu (2016). *Reshaping Medical Practice and Care with Health Information Systems* (pp. 87-131).
[www.igi-global.com/chapter/expanding-role-of-telephone-systems-in-healthcare/146005?camid=4v1a](www.igi-global.com/chapter/expanding-role-of-telephone-systems-in-healthcare/146005?camid=4v1a)