# Extracting Intention from Web Queries– Application in eHealth Personalization

G. Drosatos[1], A. Arampatzis[2] and E. Kaldoudi[1]

[1]School of Medicine, Democritus University of Thrace, 68100 Alexandroupoli, Greece
[2]Department of Electrical and Computer Engineering, Democritus University of Thrace, Xanthi, Greece

*Abstract*— **Personalizing healthcare applications requires capturing patient specific information, including medical history, health status, and mental aspects such as behaviors, intentions, and attitudes. This paper presents a privacy-friendly system to deduce patient intentions that can be used to personalized eHealth applications. In the proposed approach patient intention is deduced from web query logs via query categorization techniques. The architecture assumes a user application which conceals the user's queries from the central system, while only relevant intentions are disclosed. The paper presents a prototype implementation of the proposed architecture to extract intentions for personalizing empowerment services for the cardiorenal patient. Emphasis is placed on identifying intentions related to travel, diet and physical exercise, as these play an important role for the daily management of cardiorenal disease.**

*Keywords*— **Patient intention, patient empowerment, cardiorenal disease, privacy, personalization.**

## I. Introduction

A current challenge in health and care is to create personalized services. Personalizing eHealth applications requires capturing personal information, which most often includes medical history and current biometric readings. However, of equal importance are psychological, behavioral and cognitive aspects, for example the patients' plans, intentions, attitudes, and behavior [1].

The first two types of personal information, medical history and current medical status, are commonly harvested from personal health records and from wearable sensors respectively. Capturing information on personal plans, intentions, and other psychological and cognitive issues is rather more complicated. So far, such information is generally derived via interviews and questionnaires, a rather cumbersome process requiring human intervention. Current research however increasingly focuses on the person's web involvement as a source for information on psychological and cognitive issues [2]. In particular, web searches have shown to be a good source to reveal user's interests and intentions [3, 4]. Also, web search engines are one of the most popular uses of the web. For example, a recent USA survey [5] shows that 72% of internet users have searched online for health information of one kind or another within the past year while eight in ten online health inquiries start at a search engine.

In this paper we present a privacy friendly system to extract certain patient intentions that can be used to personalize empowerment and decision support services for the cardiorenal patient, as part of the FP7-ICT project CARRE: Personalized Patient Empowerment and Shared Decision Support for Cardiorenal Disease and Comorbidities (Grant no. 611140). Patient empowerment mainly refers to patient awareness, engagement, and control [6]. Therefore, personalization here is strongly related to the need and intention of the patient to acquire new information about a particular health concept. Additionally, in the case of the cardiorenal comorbid patient and other lifestyle related diseases, an important part of care plans and decision support involves two different lifestyle aspects, namely diet and physical activity. In this context, knowledge on patient's plans to travel is important for the personalization of diet and physical exercise, as both need to be customized to weather conditions, daily routine, and physical activity level.

## II. Intention Extraction as Query Classification

The extraction of user intention from web queries is usually seen as a categorization problem, and particularly, a classification problem, where queries are classified to predefined categories. These categories represent the user intentions. The query classification is a well-known problem as it is used to identify user search goal in order to improve search engine retrieval [7]. What makes web queries challenging is that they consist only of few words (2-3 words on average [8]) from which user interest must be extracted.

A variety of query classification approaches have been proposed. Most of them map the queries onto an external knowledge source, namely controlled vocabularies, search engine query logs [9], search engine results (documents) [10], controlled vocabularies [11, 12], Wikipedia category graph [13]. Alternative approaches built their own knowledge source out of generic, publicly available sources, such as the entire web or on purpose created representative web subset [12].

Web queries can be considered as private information [14], since they can reveal interests, political opinions, religious beliefs, sexual life, and other things a user may like to



D.A. Jaffray (ed.), *World Congress on Medical Physics and Biomedical Engineering, June 7-12, 2015, Toronto, Canada*,
IFMBE Proceedings 51, DOI: 10.1007/978-3-319-19387-8_350

keep private. All this information according to the European Data Protection Directive [15] constitutes sensitive personal data and requires special handling.

Extracting patient intention in order to personalize patient empowerment applications imposes certain additional requirements. In particular, for privacy reasons, the classification process should preferably take place at the user-side– not in a remote, 3rd party server. Thus, the classification method should be able to execute on a relatively low end personal device like a smartphone or tablet, that is, it should require minimum possible CPU, main memory, and disk space. Additionally, to account for the diversity in medical domain and goal of patient empowerment applications, the classification method should allow for updates and refinements of the query category tree and the respective training data set. To satisfy these requirements we adopt and adapt a query classification proposed by Agrawal et al [12]. The basic idea of this approach is to enrich the categories with documents from the web and classify the web queries in these documents. As shown in Fig. 1, the query classification is a two-step process.

The first step creates the training document set. The process uses a generic list of categories representative of the entire web. Each category name is then used as a search term to fetch documents that create the representative training set for this category. Our implementation uses the generic category tree from SimilarWeb [16] web analytics service, which is further extended to include additional domain specific subcategories. For the particular goals of personalizing empowerment of cardiorenal patients, the following new categories were deemed appropriate: 'drugs and therapy', 'diagnostic tests', 'medical equipment and sensors', 'exercise and physical activity' and 'dietary supplements'. This extended category list was used to create the training document data set out of the ClueWeb09 B dataset [17], an open access 50-milllion English web pages dataset created to support research on information retrieval. For each category, we collect the top-m ranking results of documents and associating them with the category label (for this implementation m = 300 as suggested in literature [12]). The ranking score of document corresponds to the probability that each document belongs to a specific category. Finally, all the collected documents for all categories are indexed with the Indri search engine [18] creating an index of categorized documents.

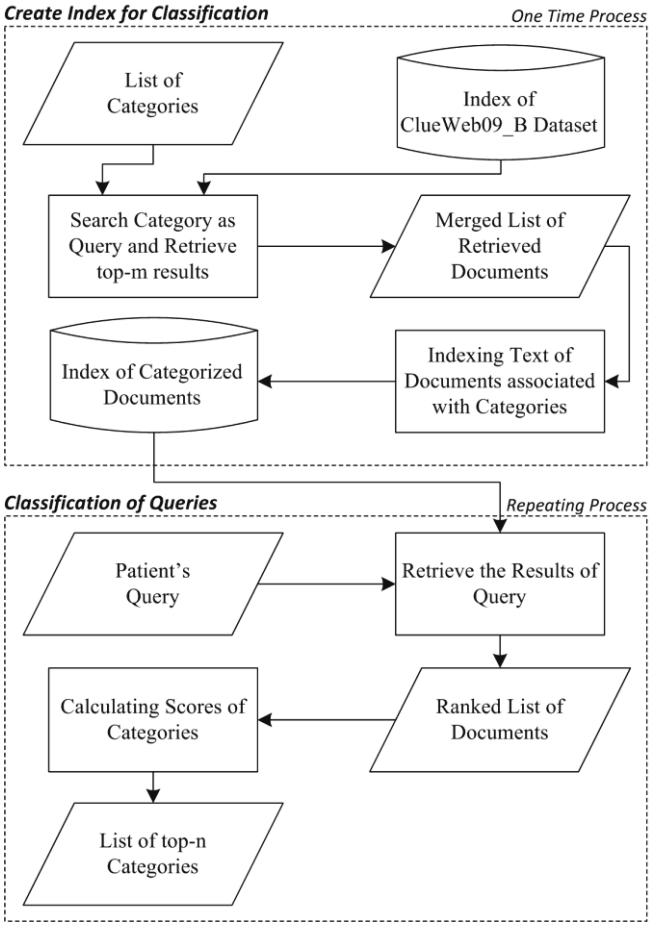


Fig. 1 The flowchart of the query classification process.

The second step performs the actual web query classification task. The patient web query is first searched in the training document set produced in the previous step. The result of the search is a ranked list of documents. The probability a query belongs to a certain category is then calculated as a function of the number of returned documents in this category weighted by their ranking scores. In particular, assume that the result of the search is a ranked list of documents $d_j$, associated with their decreasing LM (Language Model) [18] scores $s_j \in (-\infty, 0)$ with respect to the language model of the query. Note that a document may be associated with more than one category, if it is retrieved in the top-m for several categories in the previous step. For each category $c_i$, we calculate its score $S(c_i)$ as

$$S(c_i) = \sum_{j:d_j \in c_i} 2^{s_j} \quad (1)$$

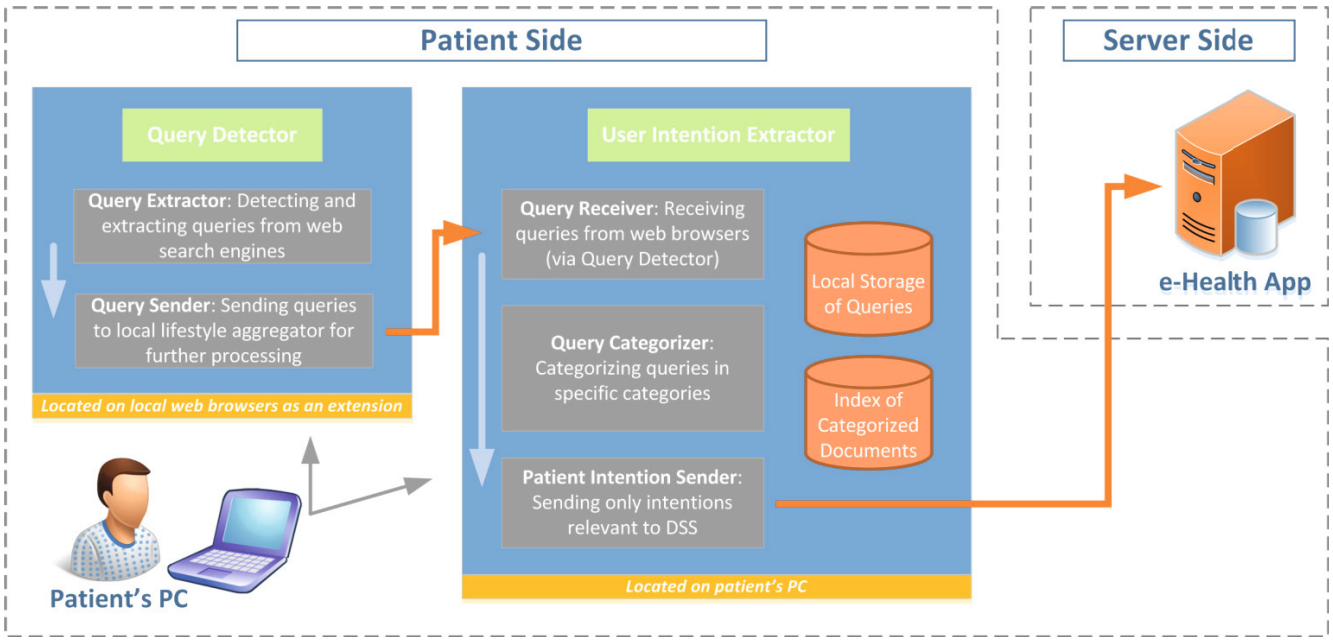


Fig. 2 The general architecture of our approach.

Then, we rank all categories in a decreasing score, and keep the top-n categories. The number of top-n categories is set to be n = 3 based on [12].

## III. Patient Intention Extraction System

In this section, we describe the proposed privacy-friendly patient intention extraction system. An overview of the architecture is presented in Fig. 2. The main components are the *Query Detector* and the *User Intention Extractor*. Both of them are located in patient-side and particularly in the personal computer of the patient.

The first functional unit of the *Query Detector* is the *Query Extractor*. This is responsible for detecting the user's queries in the web search engines of Google, Bing and Yahoo. It is triggered when the user tries to search something on the Internet. The detection of queries is achieved by parsing the URLs of the opening pages; the steps that are followed are: (1) Detect if a URL is from a search engine. (2) Parse the parameters of URL. (3) Export the query from the parameters. In the subsequent step, the *Query Sender* is responsible for forwarding the detected queries over HTTP locally (using localhost) to the *Query Receiver* of the *User Intention Extractor* for further processing.

The *User Intention Extractor* constitutes a standalone application that runs automatically when the operating system starts. It is responsible to store locally the incoming queries and categorize them in specific categories (e.g. travelling, health diseases, etc.) in order to extract the patient's intentions. The first component is the *Query Receiver*, which receives the user's queries from the local browsers' extensions and stores them to its *Local Storage*. In order to receive data from a browser extension, it is capable of handling HTTP requests. Subsequently, the *Query Categorizer* uses as input the queries from the *Local Storage* and categorizes them following the procedure described in the previous section. Finally, the *Patient Intention Sender* is responsible for exporting only the relevant intentions (health and travel) to the respective eHealth application. The communication between the *User Intention Extractor* and the remote eHealth application is encrypted (HTTPS) and an authentication mechanism (OAuth) is used to identify the patients.

The *Query Detector* is implemented in *JavaScript* as a browser extension for the Firefox v34 and the Chrome v39 (top of Fig. 3). The *User Intention Extractor* is implemented in *Java* as a standalone application and runs in system tray when the operating system starts (bottom of Fig. 3). Both implementations are platform independent, thus can run on Windows, Linux and Mac OS X.

## IV. Discussion

In this paper, we presented a privacy-friendly architecture that utilizes a patient's web queries in order to extract intentions. The patient's intentions constitute an additional

and valuable information for personalizing eHealth applications. The proposed architecture is 'loaded' at the user-side in a way that most private queries of the patient are not revealed to anyone; only the relevant intentions (i.e. health and travel issues) are released to the central system. Finally, we developed a prototype confirming the feasibility of our approach.

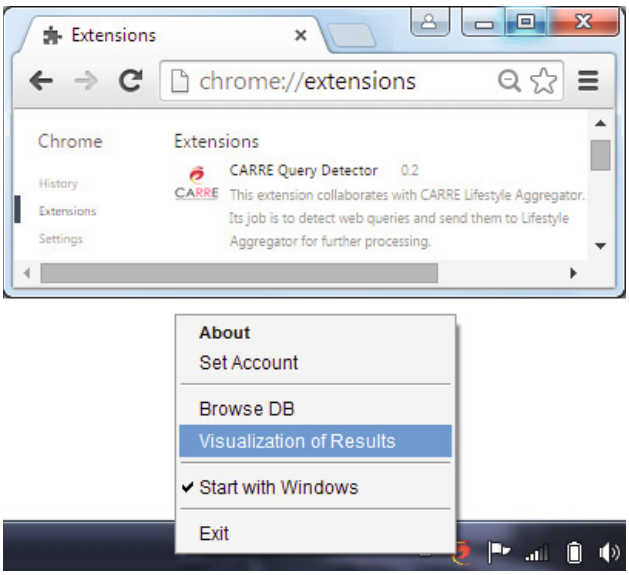


Fig. 3 A snapshot of the Query Detector (up) and User Intention Extractor (down).

Work in progress includes improvements on and evaluation of the query classification process by optimizations in the health and travelling categories related to the specific eHealth domain of cardiorenal disease.

## Acknowledgment

This work was supported by the FP7-ICT project CARRE (No. 611140), funded in part by the European Commission. G. Drosatos and A. Arampatzis were also partially supported by the project ATLAS (Advanced Tourism Planning), GSRT/CO-OPERATION/11SYN-10-1730.

## Conflict of Interest

The authors declare that they have no conflict of interest.

## References

1. Lapsley P. (2012) Involving patients in their healthcare. BMJ 345
2. Fernandez-Luque L., Karlsen R et al. (2011) Review of Extracting Information From the Social Web for Health Personalization. J Med Internet Res. 13:e15
3. Ginsberg J., Mohebbi M. H et al. (2008) Detecting influenza epidemics using search engine query data. Nature. 457:1012–1014
4. Drosatos G., Efraimidis P. S et al. (2015) Pythia: A Privacy-enhanced Personalized Contextual Suggestion System for Tourism. Tech. Rep. Euclid-TR2015-01 Athena RIC, Greece
5. The Pew Research Center. (2013) Health Online 2013. http://www.pewinternet.org/2013/01/15/health-online-2013/, Last accessed: 02 Feb. 2015
6. Kaldoudi E., Makris N (2015) Patient Empowerment as a Cognitive Process, in Proc. of HealthInf 2015: 8th International Conf. on Health Inf., Lisbon, Portugal, pp 605–610
7. Rose D. E., Levinson D (2004) Understanding User Goals in Web Search, in Proc. of the 13th International Conf. on WWW, ACM, pp 13–19
8. Arampatzis A., Kamps J (2008) A Study of Query Length, in Proc. of the 31st Annual International ACM SIGIR Conf. on Research and Development in IR, ACM, pp 811–812
9. Beitzel S. M., Jensen E. C et al. (2007) Automatic Classification of Web Queries Using Very Large Unlabeled Query Logs. ACM Trans. Inf. Syst. 25:2
10. Gabrilovich E., Broder A et al. (2009) Classifying Search Queries Using the Web As a Source of Knowledge. ACM Trans. Web. 3:5:1–5:28
11. Lovelyn Rose S., Chandran K. R (2011) Web knowledge and Wordnet based Automatic Web Query Classification. Int. J of Comp. App. 17:23-28
12. Agrawal R., Yu X et al. (2011) Enrichment and Reductionism: Two Approaches for Web Query Classification, in Neural Information Proc., Springer, 7064 of LNCS, pp 148-157
13. AlemZadeh M., Khoury R et al. (2012) Query Classification using Wikipedia's Category Graph. J of Emerg. Tech. in Web Intelligence 4
14. Arampatzis A., Efraimidis P. S et al. (2013) A query scrambler for search privacy on the internet. Inf. Retrieval 16:657-679
15. European Parliament (1995) Directive 95/46/EC, Official Journal L 281, pp 0031-0050
16. SimilarWeb (2014) List of all website categories in SimilarWeb. http://www.similarweb.com/category, Last accessed: 02 Feb. 2015
17. The Lemur Project (2009) The ClueWeb09 Dataset. http://www.lemurproject.org/clueweb09.php, Last accessed: 02 Feb. 2015
18. Strohman Metzler D., Turtle H et al. (2005) Indri: A language model based search engine for complex queries, in Proc. of the International Conf. on Intelligent Analysis

Corresponding author:

Author: George Drosatos
Institute: School of Medicine, Democritus University of Thrace
Street: University Campus, Dragana
City: Alexandroupolis
Country: Greece
Email: gdrosato@ee.duth.gr