Applied Medical Informatics
Original Research **Vol. 38, No. 2 /2016, pp: 73-80**

Towards Automatic Improvement of Patient Queries in Health Retrieval Systems

Nesrine KSENTINI*, Mohamed TMAR, Faiez GARGOURI

MIRACL Laboratory, University of Sfax, City ons, B.P. 3023, Sfax, Tunisa
E-mails: ksentini.nesrine@ieee.org; mohamedtmar@yahoo.fr; faiez.gargouri@gmail.com

* Author to whom correspondence should be addressed; Tel.: +21674862233; Fax: +21674862432

Received: April 13, 2016 / Accepted: June 27, 2016 / Published online: July 29, 2016

Abstract
With the adoption of health information technology for clinical health, e-health is becoming usual practice today. Users of this technology find it difficult to seek information relevant to their needs due to the increasing amount of the clinical and medical data on the web, and the lack of knowledge of medical jargon. In this regards, a method is described to improve user's needs by automatically adding new related terms to their queries which appear in the same context of the original query in order to improve final search results. This method is based on the assessment of semantic relationships defined by a proposed statistical method between a set of terms or keywords. Experiments were performed on CLEF-eHealth-2015 database and the obtained results show the effectiveness of our proposed method.

Keywords: e-Health; Healthcare; Semantic relationships; Medical informatics; Medical data

Introduction

With the evolution of technologies in these last years, health informatics has become an innovative application in the field of information technologies in order to improve health and healthcare [1]. It is best defined in the e-health context, an important paradigm in healthcare [2]. It is linked to the collection, analysis and circulation of data (health information) and it includes systems and services for patients, doctors and citizens. This term seems today as a "buzzword" used online (like other terms such as e-commerce, e-solutions), and applied to characterize everything related to medicine and computers.

It is a recent term derived from new electronic vocabulary applied to healthcare which necessitates the use of electronic equipment, such as computers or mobile devices, in order to communicate and to share health information. In fact, e-health will play a key role in the development and implementation of systems and processes that support the delivery of quality care to patients through clinical decisions based on evidence and delivering at the right time, relevant information to the right person.

For a good and sustainable e-health system, three attributes should be checked [3]: the data should be affordable for any health system user such as patients, employers. The second attribute is the acceptability to key constituents. In fact, patients and health professionals should accept the provided data and the third attribute is the adaptability- health system should respond adaptively to new diseases and to new scientific discoveries [3].

As a result, e-health allows patients to play a greater role in their own healthcare and facilitates the information access for institutions and health professionals firstly, and the flow of information in all aspects of their health secondly.

All that is mentioned, above, is summarized in this definition given by Eysenbach [4]: "e-health is an emerging field in the intersection of medical informatics, public health and business, referring to health services and information delivered or enhanced through the Internet and related technologies. In a broader sense, the term characterizes not only a technical development, but also a state-of-mind, a way of thinking, an attitude, and a commitment for networked, global thinking, to improve health care locally, regionally, and worldwide by using information and communication technology".

The main goal of e-health today, is to challenge data management, through information retrieval systems (IRS). This is a huge and very complex task to do and to find the relevant information to meet user's needs due to the fast evolution of medical data on the web.

These users known as patients have a narrow health literacy and may not have skills to effectively interact with the health system and engage in appropriate self-care [5].

In this paper, a method is proposed to interpret clinical and medical data on the web by defining semantic relationships between different terms in e-health context and to improve in the following step the user's query when he/she is browsing a relevant information for his/her need.

Material and Method

Used Material

Our aim is to improve the patient query (based on the user's need) when he/she browses for health information on the web. Materials used to evaluate our proposed method break down into two parts: the database, and test queries.

The employed ***database*** consists of about one million web pages (web documents) provided by the Khresmoi project and made available to the Clef-Ehealth-2015 task [6,7]. It covers a wide range of health issues done by both experts (doctors, nurses) and non experts (patients) in the health field. Web pages in the database are essentially related to health websites certified by the health foundation on the web. The provided documents or pages are in HTML (Hyper Text Markup Language) format with their URL (Uniform Resource Locators).

The ***queries set*** used to test the IRS, contains 67 queries that patients may ask when they observe symptoms and signs of a disease [6,7]. Each query of this set consists of two fields (num and query) illustrated by figure 1.

```
<topics>
<top>
<num>clef2015.test.1</num>
<query>many red marks on legs after traveling from us</query>
</top>
<top>
<num>clef2015.test.2</num>
<query>lump with blood spots on nose</query>
</top>
<top>
<num>clef2015.test.3</num>
<query>dry red and scaly feet in children</query>
</top>
   .
   .
   .
</topics>
```

Figure 1. A part of used queries (where: Num = query ID, Query =large description provided by the user (patient))

The Proposed Method

Dealing with the increasing amount of medical data on the web, the users and especially non-experts in the medical domain are faced with many difficulties to find information relevant to their needs. These difficulties are related to how the user can find the relevant documents in a wide collection and if his/her query fit well the user information need.

Indeed, queries submitted to IRS are generally in natural language and contain terms that express a symptom or a sign of a disease with which the patient was faced and as such, is trying to find more information about the disease that he/she may have.

Recent studies demonstrated that IRS fails to return the most relevant documents or web pages that meet this type of queries used by health consumers which are generally long and contain ambiguous terms.

In this context, our method proposes to automatically improve the user's need (patient query) with most related terms to the initial context provided by the user. The proposed method is based on a proposed statistical method applied to the large database in order to define semantic relationships that may exist between database keywords or terms.

The process of our method is shown in Figure 2. The figure shows the specific steps of our proposed method, the definition of semantic relationships between terms, and the automatic query expansion using the defined relations previously.

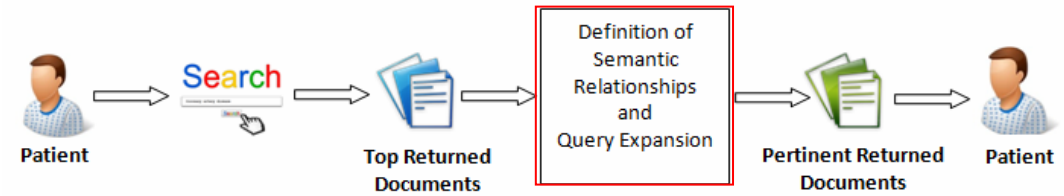

Figure 2. The process of our proposed method

In fact, when the patient submits his/her query, our information retrieval system returns the top (k) scored documents deemed as relevant. Then, it defines the semantic relationships that may exist between the terms of these documents. Once these semantic relationships are defined, they will be studied in order to expand the original query with the most related terms to query terms and to retrieve the relevant documents that meet the new query and display them to the patient.

Indeed, for each patient query, around one million web medical documents will be searched using the top 100 documents assumed as relevant using the standard platform called Terrier. Terrier is an effective open source engine, deployable for large collections of documents. It implements functionalities of indexing and retrieving documents on the web [8].

Thereafter, semantic relationships that may exist between all terms in the top 100 returned documents are defined with the following form:

$$t_i = f(t_1, t_2, \ldots, t_{i-1}, t_{i+1}, \ldots t_n) \quad (1)$$

where (n) represents the number of terms in the top 100 documents.

Least square method (LSM) [9-12] is a commonly used statistical and mathematical method for solving such problems. In fact, LSM seeks to find the relation that may exist between an explained variable (y) and an explanatory variable (x). It is a procedure to find the best regression line to (x_i ; y_i) data observed for (for i=1..n) in this way:

$$y = ax + b \quad (2)$$

whose *b* represents the residual or the error which is the disruption of the regression model.

The purpose of this method is to find the *a* values that minimize the error (Err) given by:

$$Err = \sum_{i=1}^{n} (b_i)^2 = \sum_{i=1}^{n} (y_i - ax_i)^2 \quad (3)$$

In our case, let the term t_i (for i=1..n) as the explained variable and the remaining terms in the top documents as the explanatory variables. As a result, we obtain the semantic relationships between terms at the following form:

$$t_i \approx \alpha_1 t_1 + \alpha_2 t_2 + \ldots \alpha_{i-1} t_{i-1} + \alpha_{i+1} t_{i+1} + \ldots + \alpha_n t_n + b \quad (4)$$

where α_i represent the values of relationships between the current term t_i and the remaining terms in the whole set of keywords.

To apply this statistical method (LSM) on a corpus of documents in order to obtain such results, we must represent each document with a mathematical representation. Thus, we exploit techniques and models used in the information retrieval domain.

Indeed, we use the vector space model (VSM) to represent each document by a set of keywords or terms vector. Values in these vectors represent the calculated weights of each term in the document using the TF-IDF (Term Frequency- Inverse Document Frequency) measurement. This weight is obtained by multiplying measures of (TF) and (IDF) with:

$TF(t_i)$: represents the frequency of a term (t_i) in a document

$$IDF(t_i) = \log(|D|/|d_j : t_i \in d_j|) \quad (5)$$

where

|D|: total number of documents in the corpus.

$|d_j : t_i \in d_j|$: the number of documents where the term(t_i) appears.

At the end, a "terms × documents" matrix (X) is obtained where each cell represents the weight of each term in each document. Then, we applied LSM which provides a way (see equation 5) for the matrix representation to calculate the vector A_i which contains different values of relationships α_i for each term(t_i) in a returned corpus with the remaining terms.

$\forall i = 1,..,n$

$$A_i = \left(X^{iT} * X^i \right)^{-1} * X^T * t_i \quad (6)$$

where X^i: represents the new matrix after removing the row of the current term t_i; A_i: represents the vector of relationships values (α_i) between the term (t_i)and all remaining terms.

This process of defining semantic relationships between terms with a new statistical method applied in IR field and for medical data is illustrated by the Figure 3.

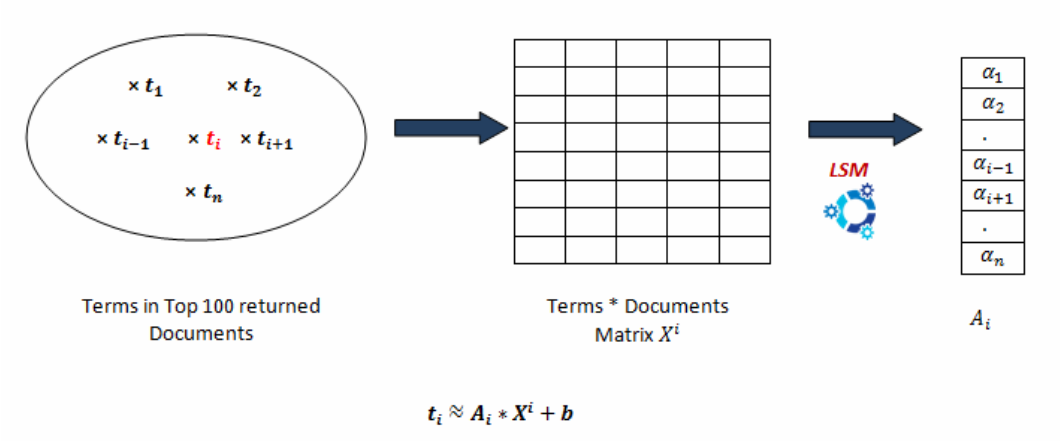

Figure 3. The process of defining semantic relationships

Indeed, for each term in the query, if it belongs to the list of terms in the top returned documents by the first pass retrieval, then we calculate relationships between this term and each term in the list.

Afterwards, weights of relationships between query terms and terms in the list are assessed. If a term is related to the initial query (the weight of relationships is above a threshold), then it will be added to the query. At the end, the new query is submitted to the system and we retrieve the most related documents.

Results and Discussion

Figure 4 presents a weighted graph of relationships between query terms (query number 7) and the most related terms in the list of terms in the top returned documents. In fact, in our case, we assume that a term is related to the query if the value of relationship is positive (α_i>0). Each edge in the graph was represented by the mean value of relationships *(mrv)* linking query terms and the term assumed as related.

$$mvr = \frac{\sum_{q}^{j=1} \alpha_{ij}\left(t_j, t_i\right)}{q} \tag{7}$$

where q: represent the number of query terms, α_{ij}: represent the value of relationship between the term (t_i) and assumed as related and the query term (t_j).

Taking as an example the relation between the term (skin) and the query number 7, the value of this relation is equal to 0.42. Indeed, this value is calculated knowing that:

α(skin, rosacea) = 0.562

α(skin, symptom) = 0.294

Thus,

$$\alpha(\text{skin, query 7}) = \frac{\alpha(\text{skin}, \textit{rosacea}) + \alpha(\text{skin}, \text{symtom})}{2} = \frac{0.562 + 0.294}{2} = 0.42$$

Weighted graph illustrated by Figure 4 show that all added terms give information about the same context of the initial query. For more generalization in order to demonstrate the effectiveness of our proposed method, Table1 presents some queries expanded by new terms.

For each expanded query, terms that are written in bold red font represent the initial query terms submitted by the patient, and the other terms represent the added terms to the original query in their root form (terms in the list are all saved in their root form after the indexation process of the database). Added terms usually express the same context of the original query which proves the effectiveness of our proposed method.

Table 1. Obtained results for some queries after query expansion method

Query ID	Initial Query	Expanded Query
7	rosacea symptoms	**rosacea symptom** prevent skin sunscreen cream vital-health-zon cheek rhinophyma oracea isotretinoin tummy parkinson conceal felderman seborrh prosacea
18	poor gait and balance with shaking	**poor gait balanc shaking** skill physic fall strength aquat anger kareus symmetri vareniclin methylphenid
34	cavity problem	**Caviti problem** mouth breastfeeding hygien floss tooth cari inlay children dentistri handpiec checkup peridex
53	Swollen legs	**Swollen leg** swell sprain clot podiatri indepth ofloxacin
66	treatment of coughs in babies	**treatment cough babi** nose sneez dismiss bout mehtachildcar epiglott hpv diphtheria diaphragm

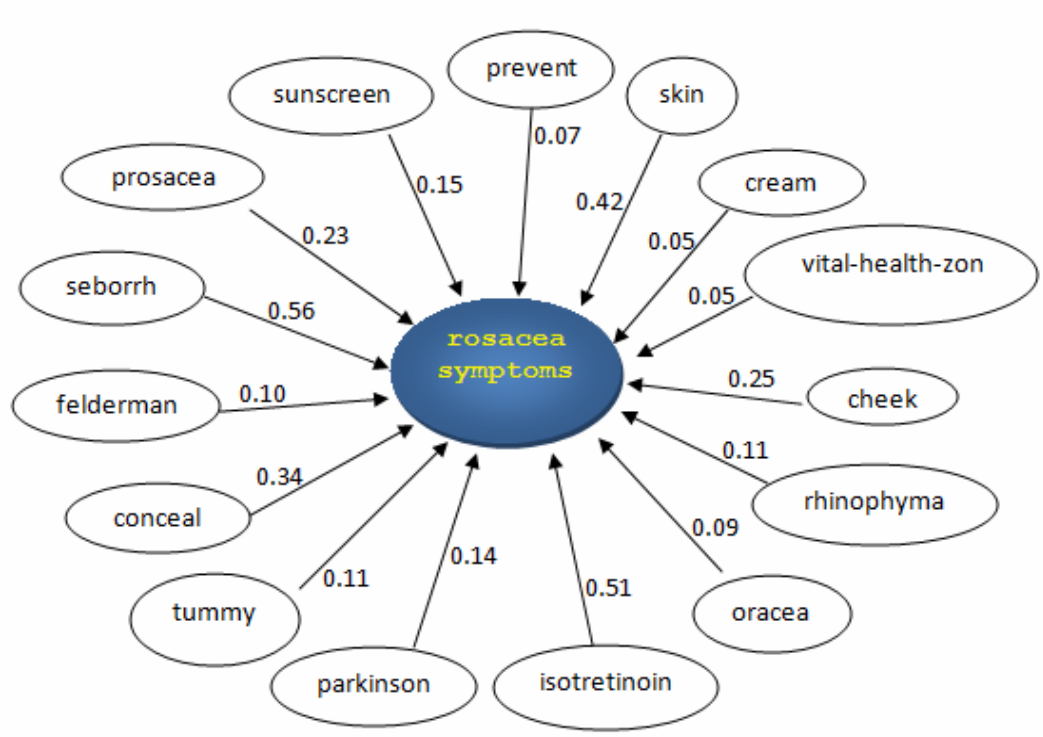

Figure 4. Graph of relationships between query number 7 and related terms

Indeed, taking for example the expanded query number 7, we explain some added terms such as:

- **rhinophyma** is a rare skin disease characterized by thickened skin and sebaceous (oil) glands are enlarged, and bumpy nose. It is sometimes also called "bulbous nose" or "Phymatous rosacea." The exact cause is unknown. However, this condition usually occurs in very severe cases of rosacea with the same symptoms of the rhinophyma disease [13,14].
- **isotretinoin**is an oral treatment already observed to be efficient in handling rosacea [15].
- **parkinson** is a disease whose risk, that a person may be affected, can be increased linked to rosacea [16]. This disease may reflect a symptom of rosacea disease.
- **felderman**: represents the name of a famous doctor in dermatology [17].

Conclusions

Obtained results show the effectiveness of our proposed method to improve, with a semantic way and through an automatic query expansion technique, the patient queries which are generally fuzzy queries. The new obtained queries are meaningful and try to help users to express their need using medical terms. Indeed, added terms are generally in the same context sought by the patient.

List of abbreviations

IRS = Information Retrieval Systems
LSM = Least Square Method
VSM = Vector Space Model

Conflict of Interest

The authors declare that they have no conflict of interest.

Acknowledgements

The authors in this paper would further like to thank the team of CLEF-ehealth-2015 competition who provided us the large database of medical web pages during our participation in this competition.

References

1. Herland M, Khoshgoftaar TM, Wald R. A review of data mining using big data in health informatics. Journal of Big Data 2014;1:2. Doi: 10.1186/2196-1115-1-2
2. Silber D. The case for eHealth. European Institute of Public Administration 2003. In European Commission's First High-level Conference on eHealth, 2003.
3. Fineberg HV. A successful and sustainable health system-how to get there from here. New England Journal of Medicine 2012;366(11):1020-7.
4. Eysenbach G. What is e-health. J. Med. Internet Res. 2001;3(2):e20. doi: 10.2196/jmir.3.2.e20.
5. Jacobs RJ, Lou JQ, Ownby RL, Caballero J. A systematic review of eHealth interventions to improve health literacy. Health Informatics Journal 2016;22(2):81-98.
6. Palotti J, Zuccon G, Goeuriot L, Kelly L, Hanbury A, Jones GJ, et al. Clef ehealth evaluation lab 2015, task 2: Retrieving information about medical symptoms. In CLEF 2015 Online Working Notes. CEUR-WS, 2015.
7. Ksentini N, Tmar M, Gargouri F. Miracl at Clef 2015: User-Centred Health Information Retrieval Task. Proceedings of the ShARe/CLEF eHealth Evaluation Lab, 2015. Available at: http://ceur-ws.org/Vol-1391/91-CR.pdf.
8. Ounis I, Amati G, Plachouras V, He B, Macdonald C, Lioma C. Terrier: A high performance and scalable information retrieval platform. In Proceedings of the OSIR Workshop, 2006, pp. 18-25.
9. Ksentini N, Tmar M, Gargouri F. Detection of semantic relationships between terms with a new statistical method. In Proceedings of the 10th International Conference on Web Information Systems and Technologies. 2014, pp. 340-3.
10. Abdi H. The method of least squares. Encyclopedia of Measurement and Statistics. CA, USA: Thousand Oaks, 2007.
11. Ksentini N, Tmar M, and Gargouri F. Controlled automatic query expansion based on a new method arisen in machine learning for detection of semantic relationships between terms. In: 15th International Conference on Intelligent Systems Design and Applications (ISDA). IEEE, 2015, pp. 134-9.
12. Ksentini N, Tmar M, Gargouri F. The Impact of Term Statistical Relationships on Rocchio's Model Parameters for Pseudo Relevance Feedback. International Journal of Computer Information Systems and Industrial Management Applications 2016;8:135-44.
13. Rohrich RJ, Griffin JR, Adams JR, William P. Rhinophyma: review and update. Plastic and Reconstructive Surgery 2002;110(3):860-9.
14. Böhm D, Schwanitz P, Stock Gissendanner S, Schmid-Ott G, Schulz W. Symptom severity and psychological sequelae in rosacea: results of a survey. Psychol Health Med. 2014;19(5):586-91.
15. Park H, Del R, James Q. Use of oral isotretinoin in the management of rosacea. Journal of Clinical & Aesthetic Dermatology 2011;4(9):54-61.
16. The JAMA Network Journals. [internet]. Rosacea linked to increased Parkinson disease risk

in Danish population study. ScienceDaily. ScienceDaily, 21 March 2016. Available from:http://www.sciencedaily.com/releases/2016/03/160321114445.htm
17. Lenora F [internet].2010. Available from: http://www.feldermandermatology.com/.