Interactive Data Visualization for eHealth Retrieval System

Nesrine Ksentini, Mohamed Tmar, and Faïez Gargouri

1 Introduction

Data visualization represents the effective presentation of information and involves a multidisciplinary communication approach. Its goal is to communicate a specific message to a user. Indeed, a visual representation of data has a main goal to communicate quantitative and qualitative information clearly and effectively through graphical means which can be static, animated or interactive [1, 2].

Selecting a color schema in data visualization process is also very important. It allows the designer to set the tone of these visualizations and try to keep a consistent representation [1].

Data visualization has been used to tackle several challenges in many disciplines such as economics, medicine, and education. As eHealth is an actual topic for today and extremely important to all practitioners, we highlight in this paper the importance of data visualization in this area. In fact, many opportunities have received attention to date for supporting people to make health document sense and for supporting them in better understanding their own illnesses and their health conditions to manage them more effectively [3, 4].

Indeed, clinical-researchers are confronted today with a huge and complex patient records based on which they must study them to make sure quality control, and discover new diseases [5].

The same applies to people as they are becoming more aware of for their own health. They need to understand their own diagnostics to improve and manage their health and to better communicate with their doctors.

N. Ksentini · M. Tmar (✉) · F. Gargouri
MIRACL Laboratory, City ons Sfax, University of Sfax, Sfax, Tunisia
e-mail: mohamed.tmar@isimsf.rnu.tn; mohamed.tmar@isims.usf.tn; faiez.gargouri@isims.usf.tn

L. Chaari (ed.), *Digital Health in Focus of Predictive, Preventive and Personalised Medicine*, Advances in Predictive, Preventive and Personalised Medicine 12,
https://doi.org/10.1007/978-3-030-49815-3_13

The body of this paper will be as follows, we start with Sect. 2 to highlight earlier works related to data visualization. The Sect. 3 sheds light on our proposed data visualization system based on LSM (Least Squares Method) which is a statistical method arisen in machine learning to find semantic relationships between a set of terms and not only between a pair of terms. The Sect. 4 gives the obtained results followed by conclusions and future works.

2 Medical Visualization Systems

In this section, we review the state of the art of data visualization systems to support users (patients and clinical researchers for example) to understand personal health information. In eHealth topic, data or information visualization is part of an overall visualization field that incorporates both information and scientific visualization, which are defined in the literature separately and considered as different [3].

Scientific visualizations represent scientific concepts such as molecules, parts of the human body, or natural phenomena, mainly in 3D [6]. The goal of these visualizations is essentially the confirmation or rejection of a particular hypothesis. Information visualization is a visualization tool to represent abstract concepts or terms. Author in [7] classifies this kind of visualization as exploratory analysis visualizations, due to their goal, help user's to find a hypothesis. The power of this tool derives from its ability to represent a large information at once, including internal relations.

In our case, we focus on information visualization and through the literature review, we can find several main researches in medical field, which investigate this kind of visualization [3].

In [8] authors proposed $AsbruView$ tool, a visualization tool developed to assist in handling treatment plans in Asbru. $AsbruView$ relies on a graphical metaphor where plans are represented as a running track which the physician 'runs' along while treating the patient [3].

$LifeLines$ proposed by [9] was one of the first tools to be used for the electronic health records representation. It was originally developed as a general-purpose visualization tool to represent personal histories that was then applied for the visualization of patient records. Our aim in this paper, is to develop a visualization tool, integrated in content based retrieval system, that presents semantic relationships between medical terms appeared in patient records and not to present only patient records data. The goal is to help users to make sense of returned documents when they use content based retrieval systems and search their needs.

3 Proposed System

In this section, we present our proposed content based retrieval system that incorporates two main steps. The first step is based on a local automatic documents-analysis to define semantic relationships between query terms and terms of the top *m* returned documents deemed relevant. The second step illustrates how to visualize these relations in a graph to help users making sense of the returned documents. The process of our search system is illustrated in (Fig. 1)

3.1 Semantic Relationships

Measuring similarity and semantic relationships between terms in a set of documents has become a primary task and plays an important role in the natural language processing (NLP) field in order to improve and to interpret search results [10]. It is the backbone of several applications, such as query expansion, disambiguation, automatic creation of thesaurus [11].

Previous approaches that study this latter idea, can be classified into three main categories [12–14]: those based on semantic knowledge (such as ontologies, data dictionaries), those based on content-based methods documents using general statistical methods [15, 16], and hybrid approaches that combine the earlier two categories.

In our case, we will adopt statistical methods to define semantic relationships between documents terms. The choice to adopt this type of method is justified first, by the independence of this process to the used language and secondly, by its ability

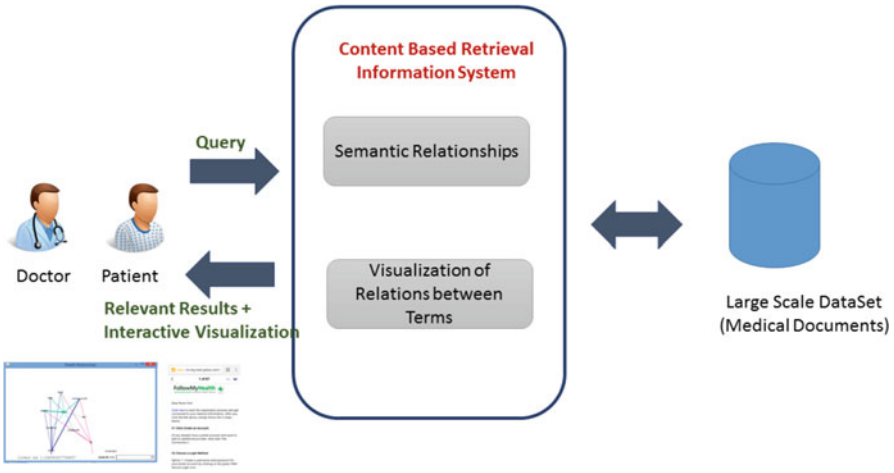

Fig. 1 Proposed content based retrieval system

to define these relations between a set of terms and not only between a pair of terms. We will apply Least Squares Method (LSM) for text analysis [17, 18] which is a method often used to define approximately relationships that may exist between many variables [17, 19–22]. Indeed, this method, known as linear regression, is the most widely used predictive model in the field of machine learning which present a particular approach to artificial intelligence [23].

Indeed, machine learning is a data analysis method which automates analytical model building process. The main idea of this method is to create algorithms that can receive input data and use statistical analysis to predict an output value.

It is an approach of artificial intelligence based on the idea that systems can learn from data and make decisions with little user intervention.

LSM tries to find the connection that may exist between an explained variable (y) and explanatory variables (x). In our case, we take a $term_j$ as an explained variable and the remaining terms in a set of documents as explanatory variables ($term_{1\cdots n}$). The goal is to find the relation between these variables as follows:

$$term_j \approx \sum_{i=1}^{j-1} (\alpha_i term_i) + \sum_{i=j+1}^{n} (\alpha_i term_i) + \epsilon \tag{1}$$

where α represent the real coefficients of the regression model and the weights of relationships between terms. ϵ represents is the associated error.

Explanatory variables are defined from the top m returned documents that meet user's needs. As a result, for each variable which is a term in our case, we will have m measures that represent the $tf-idf$ weights. To minimize calculation complexity, we study as explained variables $term_j$ only the distinctive terms of the user's query (The process of this step is illustrated by Fig. 2).

For example, when a user sends a query with three terms (t_i,t_k,t_l), our content based retrieval system retrieves the top m returned documents which will be treated with a matrix representation. Indeed, we obtain ($terms \times documents$) matrix with ($n * m$) size where n presents the number of terms in the set of returned documents.

For each query term, if it exists in the terms set of returned documents, we calculate then its relationship weight vector $A_{term_j} = (\alpha_1, \cdots, \alpha_n)$ with other n terms. Least Squares Method gives the solution to find this vector in an approximate way:

$$A_{term_j} = \left(X^{jT} \times X^j\right)^{-1} \times X^T[., j] \times T_j \tag{2}$$

Where:

T represent TF-IDF weight vector of $term_j$.
X present the TF-IDF matrix whose columns represent the keyword set and rows represent the m returned documents.
X^j is obtained by removing the column of the term t_j in matrix X.
$X^T[j, .]$ represents the transpose of the weight vector of the term t_j in all documents.

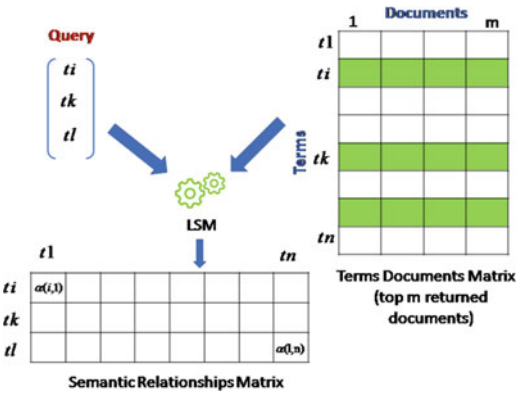

Fig. 2 Process of defining semantic relationships

At the end of this process, we obtain terms by terms matrix (see Fig. 2) (Query-terms $\times$ terms of top m returned documents matrix) which contains the relation values founded for each query term with the remainder terms.

Once the relations are defined, we study them in order to design the graph of semantic relations with the most related terms (for example terms that have positive relations with query terms).

3.2 Data Visualization

Semantic relationships defined in the previous section does not make it possible to interpret the similarity between the terms in an easy way. It is thus preferable to have a comprehensive view of these semantic relationships in order to better assimilate them.

As visualization plays a very important part in the results interpretation, we propose to visualize the defined relations in a graph which will be generated after the search process for each user's query.

The generated visual graph comprises a set of nodes and a set of edges representing respectively terms and semantic relationships between these terms. We have decided to color semantic relationships defined for each query term by a different color because visual sweeping of colors takes less time and effort.

To enhance the importance of defined relations, we have modified the color intensity and the thickness of arcs which will be proportional to the similarity value. Indeed, if the defined relationship between two terms is strong, the arc becomes thicker. In order to help users interpret results, we have used the research option of the tool Prefuse[1] which makes it possible for users to easily find a term searched in the whole graph (for example term *coronari* in Fig. 3).

[1] http://prefuse.org

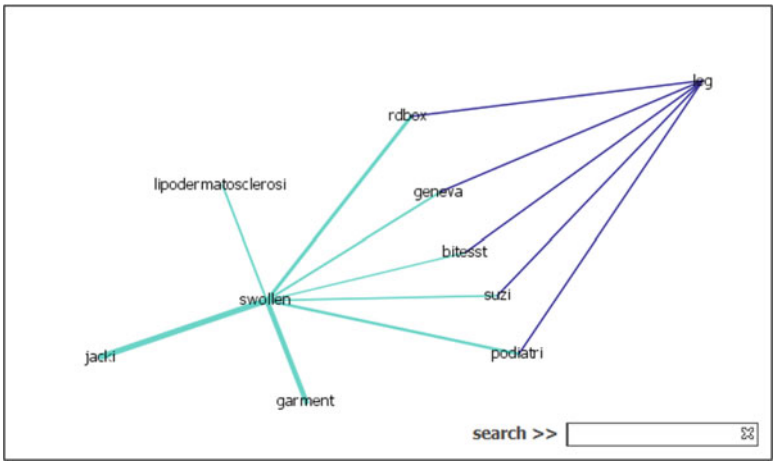

Fig. 3 Graph of Semantic Relationships: query 53 in 2015

4 Results

In order to check the performance of our proposed method, an experimental procedure was set up. This evaluation is performed on a large collection of documents provided from the CLEF company for the two successive years 2014 and 2015 [24–27].

4.1 Document Collection

The document collection is composed of a set of medical documents covering a wide set of medical topics. This collection is around of one million documents provided by the Khresmoi project [24–27] which come from different online sources such as known databases and medical sites (e.g. ClinicalTrial.gov, Genetics Home Reference, the health certified websites).

The test set in 2014 comprises 50 professional and medical queries provided by experts (clinical-researchers for example). These queries present different cases of patient diseases. In 2015, test set has 67 circumlocutory queries provided by patients when they are faced with symptoms and signs of a medical condition.

Table 1 Semantic relationships examples

Base	Query number	Some terms in relation with the query
clef2015	34	**Caviti problem** tooth cari dentistri fluoridepr dentahealth gum
clef2015	53	**Swollen leg** swell clot podiatri garment suzi lipodermatosclerosi
clef2014	1	**Coronari arteri disease** mean myocardi bypass angiograph aortic charlson cholesterol heart-attack translesion revascularis anastomos
clef2014	35	**Peptic ulcer disease** antacid food diagnos recommend gastric oesophag

4.2 LSM Results

Table 1 shows the results of defined semantic relationships for some queries. Terms that are written in bold font represent the original query terms, the other terms represent the most related terms to the original query in their root form. We notice that selected terms usually express the same context of the original query which proves the effectiveness of our proposed method.

For example we take query number 53 in 2015 and we explain some related terms like:

- podiatri: a podiatrist is a health professional who diagnoses and treats disorders of the feet.
- garment: is a pneumatic antishock garment an inflatable garment used to combat shock, stabilize fractures, promote hemostasis and increase peripheral vascular resistance.
- suzi: Extra Roomy Shoes that are cleverly designed to look slim-line but have lots of room for swollen feet.
- lipodermatosclerosi: is a skin and connective tissue disease.

Another example query 35 in 2014, related terms are in strong relation with the original query which talk about peptic ulcer. It is a disease that has long been considered chronic, defined anatomically by a loss of parietal substance not exceeding the submucosa.

4.3 Interactive Data Visualization

Interactive data visualization plays an important role in making sense process. This importance grows when we talk about the eHealth field that is a current topic that draws attention of any person which is today more aware of and take greater responsibility for his health.

The above figures show the illustrations of semantic relationships defined by the statistical method LSM. Each illustration is, as already mentioned, presented as a graph whose nodes represent the terms and the arcs represent relations between these terms.

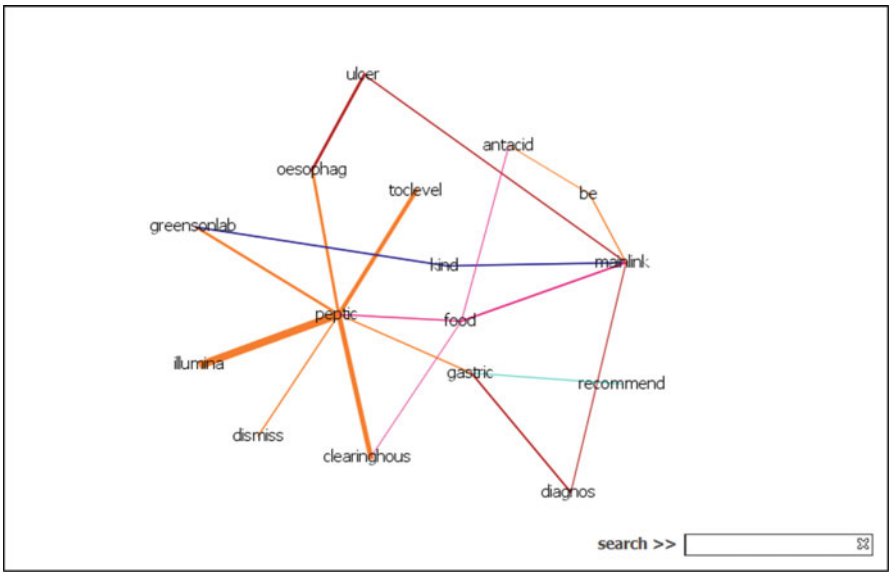

Fig. 4 Graph of Semantic Relationships: query 35 in 2014

These graphs are generated automatically and after each user's query. In fact, these types of graphs help users to understand the general context of the returned documents and to expand the knowledge on the subject later by studying the presented relationships. This study is easy to any user due to the used colors and the difference thickness edges. Indeed, if the edge between two terms is thick, the user can interpret the relation between these two terms is pertinent.

For example, in Fig. 3, edge between terms ($swollen$, $garment$) is very thick, so we conclude the strong relationships between them.

In Fig. 4, we notice the presence of several relations with the term $peptic$ that can expand user's knowledge.

In the case where we end up with a rich graph with many semantic links between the terms, the research option provided by the $Prefuse$ tool makes user's easily find a specific term. Take the example shown in Fig. 5, the user looks for the word $coronari$, our visualization tool try to seek this term, if finded this term will be colored by a pink oval.

5 Conclusion and Future Works

The goal of this work is to help users making-sense when searching medical documents. Our proposed idea is to return with the search results, a graphical representation that illustrates the semantic relationships that may exist between

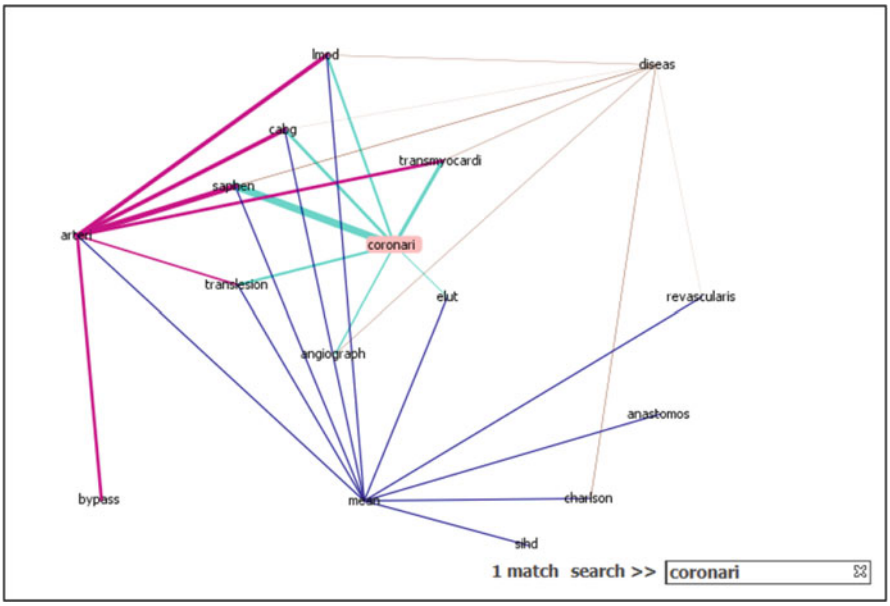

Fig. 5 Graph of Semantic Relationships: query 1 in 2014

terms in the user's query and terms of the returned documents assumed as pertinent, instead of returning only the relevant assumed documents.

Illustrated relationships in the graphs are defined by the statistical method LSM which is the most widely used predictive model in the field of machine learning that present a particular approach to artificial intelligence.

Obtained semantic relationships as well as the obtained graphs show the efficiency of the LSM method which gives significant results that help users to explore more the medical field and to ameliorate their queries with adding appropriate terms.

For future work, we try in the first time, to ask users to explain their needs (queries) based on their interpretation of provided data visualization graph. In the second time, we study the impact of query reformulation process on our content based retrieval system.

References

1. Visocky O'Grady, J., Visocky O'Grady, K.: The information design handbook. How Books, Cincinnati (2008)
2. Ksentini, N., Zarka, M., Ammar, A.B., Alimi, A.M.: Toward an assisted context based collaborative annotation. In: 2012 10th International Workshop on Content-Based Multimedia Indexing (CBMI), pp. 1–6. IEEE (2012).
3. Faisal, S., Blandford, A., Potts, H.W.: Making sense of personal health information: challenges for information visualization. Health Inform. J. **19**(3), 198–217 (2013)

4. Ksentini, N., Tmar, M., Gargouri, F.: Towards automatic improvement of patient queries in health retrieval systems. Appl. Med. Inform. **38**(2), 73–80 (2016)
5. Rind, A., Wang, T.D., Aigner, W., Miksch, S., Wongsuphasawat, K., Plaisant, C., Shneiderman, B.: Interactive information visualization to explore and query electronic health records. Found. Trends®Human-Comput. Interact. **5**(3), 207–298 (2013)
6. Card, S.K., Mackinlay, J.D., Shneiderman, B. (eds.). Readings in Information Visualization: Using Vision to Think. Morgan Kaufmann, San Francisco (1999)
7. Keim, D.A.: Visual exploration of large data sets. Commun. ACM **44**(8), 38–44 (2001)
8. Kosara, R., Miksch, S.: Metaphors of movement: a visualization and user interface for time-oriented, skeletal plans. Artif. Intell. Med. **22**(2), 111–131 (2001)
9. Plaisant, C., Mushlin, R., Snyder, A., Li, J., Heller, D., Shneiderman, B.: LifeLines: using visualization to enhance navigation and analysis of patient records. In: The Craft of Information Visualization, pp. 308–312 (2003)
10. Agirre, E., Alfonseca, E., Hall, K., Kravalova, J., Pasca, M., Soroa, A.: A study on similarity and relatedness using distributional and wordnet-based approaches. In: Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pp. 19–27. Association for Computational Linguistics (2009)
11. Terra, E., Clarke, C.L.: Frequency estimates for statistical word similarity measures. In: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, vol. 1, pp. 165–172. Association for Computational Linguistics (2003)
12. Agirre, E., Cuadros, M., Rigau, G., Soroa, A.: Exploring knowledge bases for similarity. In: LREC (2010)
13. Hassan, S., Mihalcea, R.: Semantic relatedness using salient semantic analysis. In: AAAI (2011)
14. Santus, E., Chiu, T.S., Lu, Q., Lenci, A., Huang, C.R.: Unsupervised measure of word similarity: how to outperform co-occurrence and vector cosine in VSMs. arXiv preprint:1603.09054 (2016)
15. Sahami, M., Heilman, T.D.: A web-based kernel function for measuring the similarity of short text snippets. In: Proceedings of the 15th International Conference on World Wide Web, pp. 377–386. ACM (2006)
16. Ruiz-Casado, M., Alfonseca, E., Castells, P.: Automatic assignment of wikipedia encyclopedic entries to wordnet synsets. In: International Atlantic Web Intelligence Conference, pp.380–386. Springer, Berlin/Heidelberg (2005)
17. Ksentini, N., Tmar, M., Gargouri, F.: Detection of semantic relationships between terms with a new statistical method. In: WEBIST (2), pp. 340–343 (2014)
18. Ksentini, N., Tmar, M., Gargouri, F.: Towards a contextual and semantic information retrieval system based on non-negative matrix factorization technique. In: International Conference on Intelligent Systems Design and Applications, pp. 892–902. Springer, Cham (2017)
19. Abdi, H.: The method of least squares. In: Salkind, N.J., Rasmussen, K. (eds.) Encyclopedia of Measurement and Statistics. SAGE Publications, Thousand Oaks (2007)
20. Miller, S.J.: The method of least squares. Mathematics Department Brown University, pp. 1–7 (2006)
21. Ksentini, N., Tmar, M., Gargouri, F.: Controlled automatic query expansion based on a new method arisen in machine learning for detection of semantic relationships between terms. In: 2015 15th International Conference on Intelligent Systems Design and Applications (ISDA), pp. 134–139. IEEE (2015)
22. Ksentini, N., Tmar, M., Gargouri, F.: The impact of term statistical relationships on Rocchio's model parameters for pseudo relevance feedback. Int. J. Comput. Inf. Syst. Ind. Manag. Appl. **8**, 135–44 (2016)
23. Huang, G.B., Wang, D.H., Lan, Y.: Extreme learning machines: a survey. Int. J. Mach. Learn. Cybern. **2**(2), 107–122 (2011)

24. Goeuriot, L., Kelly, L., Li, W., Palotti, J., Pecina, P., Zuccon, G., Mueller, H.: Share/clef ehealth evaluation lab 2014, task 3: user-centred health information retrieval. In: Proceedings of CLEF (2014)
25. Ksentini, N., Tmar, M., Gargouri, F.: Miracl at CLEF 2014: eHealth information retrieval task. In: Proceedings of the ShARe/CLEF eHealth Evaluation Lab (2014)
26. Palotti, J.R., Zuccon, G., Goeuriot, L., Kelly, L., Hanbury, A., Jones, G.J., Pecina, P.: CLEF eHealth Evaluation Lab 2015, Task 2: Retrieving information about medical symptoms. In: CLEF (Working Notes) (2015)
27. Ksentini, N., Tmar, M., Boughanem, M., Gargouri, F.: Miracl at Clef 2015: usercentred health information retrieval task. In: CLEF (Working Notes) (2015)