Semantic Doctor Assistant: An Ontology-based Disease Classification in Biomedicine

Yaman Kannot
Faculty of Engineering and Technology
Arab Academy for Science, Technology and Maritime Transport

Mohamed Kholif
College of Computing and Information Technology
Arab Academy for Science, Technology and Maritime Transport

Amani A. Saad
Faculty of Engineering and Technology
Arab Academy for Science, Technology and Maritime Transport

ABSTRACT
Human disease data is a cornerstone of biomedical research for diseases' classification and recommended treatments so; there is a significant need for a standardized representation of human diseases and an efficient algorithm for retrieving information from it. The Semantic Doctor Assistant (SDA) has been designed to help doctors to find proper information about a specific disease using semantic web technology rather than other simple keyword-based search. A preliminary usability study has been done to evaluate the system by measuring user's satisfaction through a statistical analysis of surveys. This study would measure the relevance of the information retrieved for each search query and how the system is important in the field of medicine and how it will help academic doctors in their research and non-academic doctors in their work.

Keywords
Semantic Web, Biomedicine, Classification, Spreading Activation

1. INTRODUCTION
Originally computers were used for computing numerical calculations and perform tasks without any sort of intelligence or semantics therefore there is a need to make computers as intelligent as human [1]. The shortcoming of non-Semantic applications is represented by the statement "lack of semantics" especially when the talk is about information retrieval. The Semantic Web can be defined as "the extension of the World Wide Web which is characterized by the association of machine-accessible formal semantics with more traditional Web content". The Semantic Web's goal is to improve the interoperability and increase automation in processing web-based information systems [2]. The Semantic Web brings structure to the meaningful content of Web pages making software agents can carry out complicated tasks instead of users [3] using a semantic knowledge base called Ontology. Ontologies is a repository in which information are organized and used in artificial intelligence and the Semantic Web applications [4]. In general, ontologies can be used beneficially in enterprise applications [5]. Within health informatics, ontology is a formal description of a health-related domain. The use of ontologies in medicine is mainly focused on the representation and organization of medical terminologies. There is need for a solution to help doctors to find proper information about a specific disease with more accurate results rather than other simple keyword-based search.

This paper is organized as follows. Section 2 describes the process of constructing ontology for the proposed solution while Section 3 explains the general framework of SDA system and system evaluation is reported in Section 4. In Section 5, conclusions and future work has been presented.

2. CONSTRUCTING ONTOLOGY FOR SDA
Human disease data is a cornerstone of biomedical research for identifying drug targets, connecting genetic variations to phenotypes, understanding molecular pathways relevant to novel treatments and coupling clinical care and biomedical research [6,7] so there is a significant need for a standardized representation of human disease to map disease concepts across resources, to connect gene variation to phenotypes and drug targets and to support development of computational tools that will first, robust data analysis and integration [8,9]. This phase secondly, divided into two sub phases, Finding and studying well-formed diseases ontology and modifying and updating the ontology to fulfill the desired needs.

2.1 Phase 1: Finding and Studying Well-Formed Diseases Ontology
The raw ontology (the original one) is a disease ontology (DO) [10] which has been developed as a standardized ontology for human disease with the purpose of providing the biomedical community with sustainable descriptions of human disease terms, and related medical vocabulary disease concepts. Great improvements to the DO have been made since 2012 including: content DO has had 192 revisions, including the addition of 760 concepts and 32% of all the terminology now includes definitions, and improved data structure [10]. The current version of the DO website [11] (version 1.0) provides a comprehensive resource to perform full-text searching on the DO as well as exploring and visualizing relationships between terms.

2.2 Phase 2: Modifying and updating the ontology
The ontology found in phase 1 is modified by making the following modifications:

- Adding treatments to each disease.
- Adding German scientific synonym for each disease.
- Adding synonyms for diseases and drugs.
- Adding symptoms.

3. METHODOLOGY
The research methodology is divided into two phases: Semantic search and Disease classification. In semantic search phase, semantically matched diseases have been

retrieved from the ontology even when typing one of its synonyms including German synonyms. The result set consists of semantically matched diseases each of which has its treatments, synonyms, and definition. In Semantic Matching, the proposed WAFAA algorithm is used in search process using concepts that are semantically related to query concepts. It is assumed that, when a user is searching for a concept, he/she is also interested in synonyms of that concept. For example, the synonyms of the concept "heart failure" are "Weak heart", "Cardiac Failure" and "CHF" therefore; diseases describing these concepts should be retrieved as well. The main objective of this phase is to identify the eligible terms in the query that can be annotated. Each concept of the ontology can have a set of predicates and objects [P, O]. Every resource in the ontology is compared with the terms in the query (via iteration with analysis component) and in case of (count >0), the term will be marked with reference URI of its equivalent resource in the ontology. Figure 1 shows the matching process in details.

Input: *Ontology O, String q*
Output: Disease Set DS
Integer i=0;
For each (Word win *q*) do
Count=0;
For each (resource *R* in *O*) do
Count = search_word_in_ontology(*R*, w);
Count+= search_word_in_ontology_using_synonyms(*R*, w);
If (count>0) then
DS +=Reasoner.getHierarchies(O,w);
Else
Remove w from q;
End if
i++;
End for
End for

Fig 1: Algorithm 1: WAFAA algorithm

In Disease classification process, a disease is classified by using Spreading Activation (SA) on ontology graph. Spreading activation is a method for searching associative networks, neural networks, or semantic networks. The result of classification may consist of a set of different classes as shown in figure 2. For instance, "Hypertensive heart disease" is classified as a hypertension disease and a heart disease.

Fig 2: Hypertensive heart disease is classified as a hypertension disease and a heart disease.

The proposed approach concerned to search and classify user's query using ontology-based similarity by using SA algorithm. To extract fully relevant information from ontology, a natural language processing technique to measure the semantic similarity under the biomedical domain is required. Semantic similarity refers to human judgments of the degrees of relatedness between a given pairs of concepts [12].

There are two main categories: ontology-based and corpus-based. The first class of the techniques is to measure the semantic similarity of the two concepts by calculating the distance between the concept nodes in an ontology tree or hierarchy [13], [14]. Figure 3 shows the proposed architecture for SDA. The system automatically processes user query for ontology-based searching in the following steps:

1. Users submit the query to the system using the user interface.
2. NLP processor performs the following steps:
   (a) Text Tokenization.
   (b) Identification all unique words.
   (c) Removal of stop words: (the, of, and, to, ….).
   (d) Word stemming (porter stemmer): find out the root/stem of a word.
3. Formulate a semantic query :
   (a) The system then convert it to semantic query using SPARQL [15] and executes it against the underlying ontology which comprises a set of diseases with semantic links and a set of rules defining ontology axiomatic semantics.
   (b) By incorporating these rules, implicit information about diseases related to user query is retrieved.
4. Semantic reasoner**:** Jena [16] reasoner is used for making knowledge discovery. Jena provides the reasoner interface for supporting inference engines.

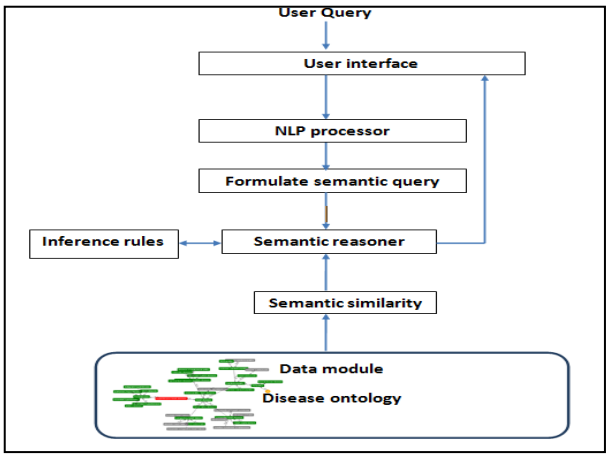

Fig 3: The proposed architecture for SDA.

4. SYSTEM EVALUATION

In order to get expressive results, statistical measures were made to show the satisfaction of doctors after using the system and the quality of the results they founded. The evaluation was carried out in two stages. In the first stage (survey design and run), the survey which is used for the evaluation was design. In this stage, the survey questions was selected carefully to evaluate different feature of the system including disease classification, synonym search, agree with information result and founding needed information from the first hit.

After that, the survey was given to 20 doctors to evaluate the system. In the second stage (statistical analysis), a statistical analysis methods have been performed on the 20 survey forms. In this section, the evaluation of the Semantic Doctor Assistant prototype is handled. The actual evaluation was done by asking 20 volunteers to act as end-users of the system. The persons involved were academic and no-

academic doctors and told to ask the system at least ten queries. Five basic example queries (in table 1) were displayed to the volunteers, serving as a starting point in order to get familiar with the system and how to phrase queries. Queries in the experiments were designed to test various search features including searching for a disease by its popular name or one of its synonyms (exact synonym or narrow synonym).

Table 1: Sample of search queries

Query #	Query text
1	Congestive Heart Failure
2	Essential Hypertension
3	Hyperuricemia
4	Toxic Shock Syndrome
5	Eumycotic Mycetoma

4.1 Statistical Analysis

The Overall evaluation for all doctors is 81% the calculated using the individual evaluation for each doctor shown in figure 4. The highest question satisfaction is 90 % (18/20) for the questions "Do you like to use SDA in your work?", "Is SDA easy to use?" and "I would be completely happy to see this system again".

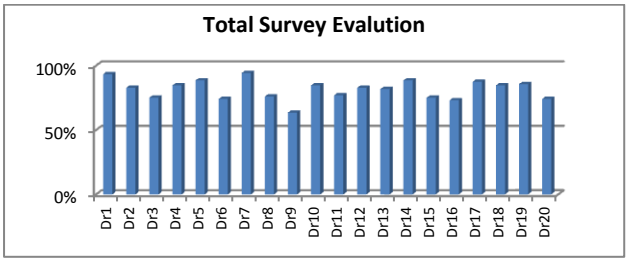

Fig 4: Individual evaluation for each doctor

User satisfaction for the questions "Do you like to use SDA in your work?" and "Did SDA found a solution of synonyms?", in Yes/No questions are shown in figure 5and figure 6 respectively. User satisfaction for the questions "How often does the system provide sufficient information?" and "How much you agree with the results compared to traditional search tools in how much questions are shown in figure 7 and figure 8 respectively.

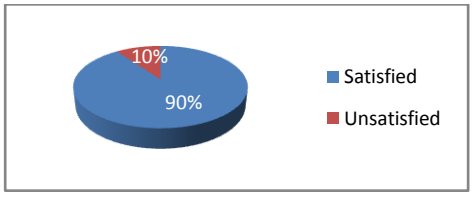

Fig 5: User's satisfaction for the question "Do you like to use SDA in your work?"

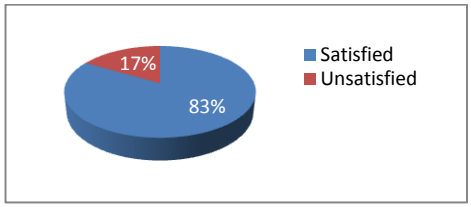

Fig 6: User's satisfaction for the question "Did SDA found a solution of synonyms?"

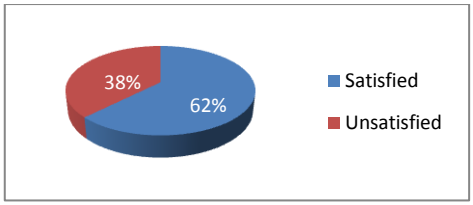

Fig 7: User's satisfaction for the question "How often does the system provide sufficient information?"

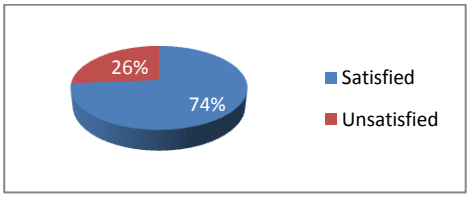

Fig 8: User's satisfaction for the question "How much you agree with the results compared to traditional search tools?"

4.2 Testing Results

The performance of the extraction process has been evaluated by using Precision, Recall and F-Measure metrics. Precision is the percentage of correctly recognized information from the total number recognized information, Recall is the percentage of information in the reference set that were recognize and F-measure is a harmonic mean of precision and recall given by:

$$Precision = \frac{number\ of\ relevant\ items\ retrieved}{number\ of\ retrieved\ items}$$

$$Recall = \frac{number\ of\ relevant\ items\ retrieved}{number\ of\ relevant\ items}$$

$$F - measure = \frac{2 \times recal \times precision}{recall + precision}$$

Figure 9 presents the average precision, recall, and F-measures of the SDA. The matching algorithm allows the system to retrieve eligible diseases.

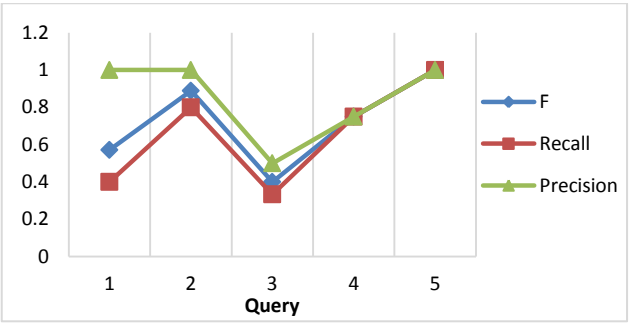

Fig 9: Precision, recall and F-measures of SDA on 5 queries from the query set.

5. CONCLUSION

This paper presents the Semantic Doctor Assistant (SDA), a full open-source semantic web prototype for find proper information about a specific disease using semantic web technology rather than other simple keyword-based search. The proposed system provides a novel method for query statement classification by parsing an ontology graph using spreading activation-based algorithm. Preliminary

experimental results show that the proposed model improves the search results of keyword-based search engines in most medical web-based system that lack the semantics of query's keywords. This avoids many irrelevant results and save time to search for the needed information. The conducted experiments on the query set have showed that the proposed system and the modified version of the OD ontology exploitation improves the search quality in terms of the precision, recall, and F- measures. For future work, it is planned to combine more ontologies to increase the relation coverage and researching methods to better recognize relations in a query.

6. REFERENCES

[1] Antoniou, Grigoris, and Frank Van Harmelen. A semantic web primer. MIT press, 2004.

[2] Payne, Terry R., and Ora Lassila. "Semantic web services." IEEE Intelligent Systems 19.1 (2004): 14-15.

[3] Berners-Lee, Tim, James Hendler, and OraLassila. "The semantic web."Scientificamerican 284.5 (2001): 28-37.

[4] Shadbolt, Nigel, Wendy Hall, and Tim Berners-Lee. "The semantic web revisited." Intelligent Systems, IEEE 21.3 (2006): 96-101.

[5] Daniel Oberle, How ontologies benefit enterprise applications, Semantic Web Journal, IOS Press, 2013. DOI 10.3233/SW-130114.

[6] K¨ ohler,S., Doelken,S.C., Mungall,C.J., Bauer,S., Firth,H.V., Bailleul-Forestier,I., Black,G.C., Brown,D.L., Brudno,M., Campbell,J. et al. (2014) The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data. Nucleic Acids Res., 4, D966–D974.

[7] Croft,D., Mundo,A.F., Haw,R., Milacic,M., Weiser,J., Wu,G.,Caudy,M., Garapati,P., Gillespie,M., Maulik,R., Kamdar,M.R. et al.(2014) The Reactome pathway knowledgebase. Nucleic Acids Res.,42, D472–D477.

[8] Robert M. Colomb, Ontology and the Semantic Web, Volume 156,2007.

[9] LePendu,P.,Musen,M.A. and Shah,N.H. (2011) Enabling enrichment analysis with the Human Disease Ontology. J. Biomed. Inform., 44, S31–S38.

[10] Kibbe, Warren A., et al. "Disease Ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data." Nucleic acids research (2014).

[11] Schriml,L.M., Arze,C., Nadendla,S., Chang,Y.W., Mazaitis,M., Felix,V., Feng,G. and Kibbe,W.A. (2012) Disease Ontology: a backbone for disease semantic integration. Nucleic Acids Res., 40, D940–D946.

[12] T. Pedersen, S. Pakhomov, and S. Patwardhan, "Measures of semantic similarity and relatedness in the medical domain," J. Biomed. Inf., vol. 40, no. 3, pp. 288–299, 2007.

[13] D. L. McGuinness, R. Fikes, J. Rice, and S. Wilder, "An environment for merging and testing large ontologies," in Proc. 7th Int. Conf. Principles Knowl. Represent. Reason., Breckenridge, CO, USA, Apr. 12–15, 2000, pp. 483–493.

[14] D. L. McGuinness, R. Fikes, J. Rice, and S. Wilder, "The chimaera ontology environment," in Proc. 17th Nat. Conf. Artif. Intell., Austin, TX, USA, Jul. 30, 2000, pp. 1123–1124.

[15] E. Prud'hommeaux and A. Seaborne, "SPARQL Query Language for RDF", URL: http://www.w3.org/TR/rdf-sparql-query/, [November 2014].

[16] Jena, "A Semantic Web Framework for Java", URL: http://jena.sourceforge.net [December 2014].