Prototype System for Multiple Sources Multiple Search Techniques Prediction

Algorithm and architecture

Marika Apostolova Trpkovska
South East European University, SEEU
Tetovo, FYROM
m.apostolova@seeu.edu.mk

Betim Cico
South East European University, SEEU
Tetovo, FYROM
b.cico@seeu.edu.mk

***Abstract*—The future of health care may be in “predictive health” that emphasizes prediction instead of diagnosis. Nowadays, the researchers are mining the data provided in social networks, aiming in prediction of diverse phenomena like social, political, medical, etc. The first part of the paper outlines the e-Health revolution phenomena. Next, we are focused on proposing a searching algorithm for predicting children general diseases in FYROM. The prediction task of the health related issues of specific people from noisy data is taken into consideration. We offer a model that can predict children general diseases with high percentage precision and good semantic recall on the basis of special designed ontology and social ties with other people, as revealed by their posts in social networks which is advised to be used by young mothers in our country. Also the architecture for ontology to database conversion is suggested and ready for implementation.**

Keywords—children disease prediction algorithm; ontology/database conversion architecture; semantic web; social network

I. THE e-HEALTH REVOLUTION

Worldwide Health Care organizations face considerable challenge to generate more suitable, efficient and resourceful tools for promoting health and providing care. The Internet itself has offered chances to face the upcoming challenges. The Internet has competency to increase the access to person health care and to enhance self-management skills. This is due to the fact that technologies based on web are comprehensive and encompass possibilities for interactivity related to information adaptation specific to the individual.

The increased possibilities of supporting health care through the use of Internet technology have led to establishing the “e-Health” concept or “electronic health”. This concept refers to all kinds of IT technology used to support a wide range of health care needs and promote the well-being. Due to the fact that e-Health has a broad scope of meaning, we have found it difficult to define the concept itself. In 2001 Eysenbach defined e-Health as “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 is comprised of not only the technical development, but also the state-of-mind, the way of thinking, the attitude, and the commitment for networked, global thinking, improving healthcare locally, regionally, and worldwide by using information and communication technology.” [1].

Based on literature review, e-Health covers a wide range of technologies, including Internet technologies, for example:

- Informational web sites
- Collaborative health communication applications (online consultation, online communities, online health decision-support programs)
- Online portals for health care
- Electronic health records (EHR) evolving concept

It also revolutionizes global health programs based on mobile communications technology, and other innovative technologies such as virtual reality (VR) and gaming technology solutions and their application in the field of healthcare (e.g. Computer Assisted Rehabilitation Environment System CAREN allows a therapist to place the patient into a virtual environment to help diagnose medical disorders like Parkinson's Disease as well as other neurological abnormalities); home automation (also called domotics); smart sensor or remote monitoring technology platforms; robots and automated systems are being deployed to assist people or to perform surgery [2].

e-Health offers possibilities to strengthen the healthcare system by delivering available and affordable high quality health care. Courage et al. point out that the e-Health model of care has potential to increase access to the health care by making available healthcare service delivery at all times, in all places, in many forms and for everyone [3]. It allows patients to receive care at whatever time and way in which they need it. This implies that the health care system must be open at all times, and access to care should be delivered over the Internet and by mobile phone in addition to head-on visits. e-Health spreads the possibility of health care outside its conservative boundaries by decreasing the limitations on traditional health care service deliverance.

Internet supports various groups, by enabling social networking for community supporting isolated individuals [4]. e-Health similarly offers opportunities to upturn effectiveness in health care sector in so cut costs [5].

As specified in the above definition for e-Health, familiarized technology involves an innovative approach of thinking about how to deliver health care that is supported by Internet technology. Patients having access to health care and communicating with other patients and caregivers about their diseases, symptoms, signs and treatments, now can use the Internet technology. This for sure will change the traditional health care delivery practice. So the e-Health can be seen as the promoter for changing the individual health care.

Nevertheless, the vital challenge of e-Health is to encourage patient-centered care. This is due to its ability to provide care that is open to individual needs, preferences and values. The effective use of information and communication technology (ICT) in health care opens up new avenues for patient-centered health information, care and services. It is challenging to use this new technology in health care and to develop applications and services for disease predictions. So, the future of health care may be in “predictive health” that emphasizes prediction instead of diagnosis [6]. It is in this domain that the Semantic Web technologies can help in realizing the goal.

Nowadays the health care users are exhausted from wasting time, money, long waiting for doctor appointments. They are also stressed with inconvenient visits scheduling and etc. The new e-Health users are running for expediency, control and choice. The traditional ways of health care delivery are changed by the shift from a role in which the actual patient is the passive receiver of health care services to an active role in which the patient is informed, has different choices, and is involved in the decision-making process.

II. Research Introduction

As discussed above, people are increasingly addicted to new modern technologies. The use of computers and the Internet are becoming a common place for people. The search engines turn out to be friends that you can ask for advice. But, it is questionable if they are trustworthy, if they are able to respond and if the result that they offer is genuine. This is particularly important when the search is related to a health issue. Often people want to search for some change in their health (symptom). Online discussion forums and social networks are enriched with content in the field of medicine through the exchange of information between the users. The Internet itself has grown into a jungle of data that an average user founds very difficult to get along with. The need to classify and organize the data grows with their increased quantity.

In addition to the standard Web of documents, World Wide Web Consortium (W3C) is helping to construct a technology stack to support a Web of data (data found in databases). The initial goal of the Web of data is to facilitate computers to do more practical work and to expand systems that can sustain trusted interactions over the global network. The Semantic Web (Web of linked data) and its technologies enable people to create data warehouses on the Web, construct vocabularies and define rules for data handling. Linked data are empowered by technologies such as RDF, SPARQL, OWL, and SKOS [7].

The Semantic Web paradigm is designed in a manner to let users make explicit statements about any data resource, and maintain them in an open and distributed manner. Several known standards like the Resource Description Framework (RDF) and Web Ontology Language (OWL) have been developed to understand the Semantic Web layer cake [8].

The development of Web 2.0 and its enrichment in Web 3.0 has resulted in generating an immense amount of blog repositories, review sites and online web discussion forums. In these sort of online discussions, people express their opinions, exchange knowledge and beliefs, give advice and criticize products and ideas. Tracking opinions on particular subject matters allows identification of user expectations and necessities. For example, peoples’ feelings about certain health decisions or reactions to particular experiences [9].

Regardless of time and place, we are witnesses of an explosion of social media usage. Online popularity has developed that spotlight equally in their individual and professional lives. Online groups that focus on every possible area of concern like nutrition, sport life, music, movies, maternity, health issues, etc have been created. As projected presently, there are over 900 social media sites on the Internet. Some of the more well known and liked platforms are Facebook, Twitter, Google Plus, LinkedIn and YouTube network [10].

A great part of population is using social media sites in some form or another. Considering the increase in the use of social media sites, there is a noteworthy amount of data that is being produced. Therefore, the people are not only joining social media sites, but they are also spending time being engaged in social media and generating an important quantity of content. Based on this, the parties become aware and are attempting to strength the power of social media to help people succeed in their requests.

Related to our effort, latest works have demonstrated that this social network data can be used to predict various phenomena including forecasting box-office revenues for movies [11], political elections [12], flu epidemics [13] and others. Other researches are focused on predicting collective properties of the popularity from blogs activities. For instance, one could try to predict whether someone will go to a movie or vote for exact candidate based on social network data. Our work concentrates on percentage prediction of children general diseases based on entered symptoms, ontology search, social network posts, social network profiles, demographic data and more.

III. The Algorithm and Architecture

The quality of medical health care as well as the increase of the efficiency of clinical practices can be significantly improved by incorporating information technology in the medical procedures which are routinely used in the medical environment.

From our preliminary review of other related works, we have found different architectures and models proposed for medical related predictions. Abhijit V. Kshirsagar at el. suggest a model that can be used to guide population-level prevention efforts and to initiate discussions between practitioners and patients about possible risk for getting kidney disease [14]. Further on, Adam Sadilek and his team suggest a scalable probabilistic model that demonstrates that the health of a person can be accurately inferred from his/her location and social interactions observed via social media [15]. Next, Chandra Sheka's work deals with an improved algorithm for prediction of heart diseases using case based reasoning technique on non-binary datasets [16]. Darcy A. Davis at el. predict individual disease risk based on medical history using a collaborative assessment and recommendation engine [17]. Steinhaeuser and Chawla use a hybrid technique based on collaborative filtering and nearest neighbor classification [18]. The similarity between two patients is computed with the Jaccard coefficient [19], which is the normalization of common diseases that two patients have, with respect to their union. What was found with this works was the similarity in the field of interest. But, the complexity and one variable concentration in their proposed models for prediction have to be stressed.

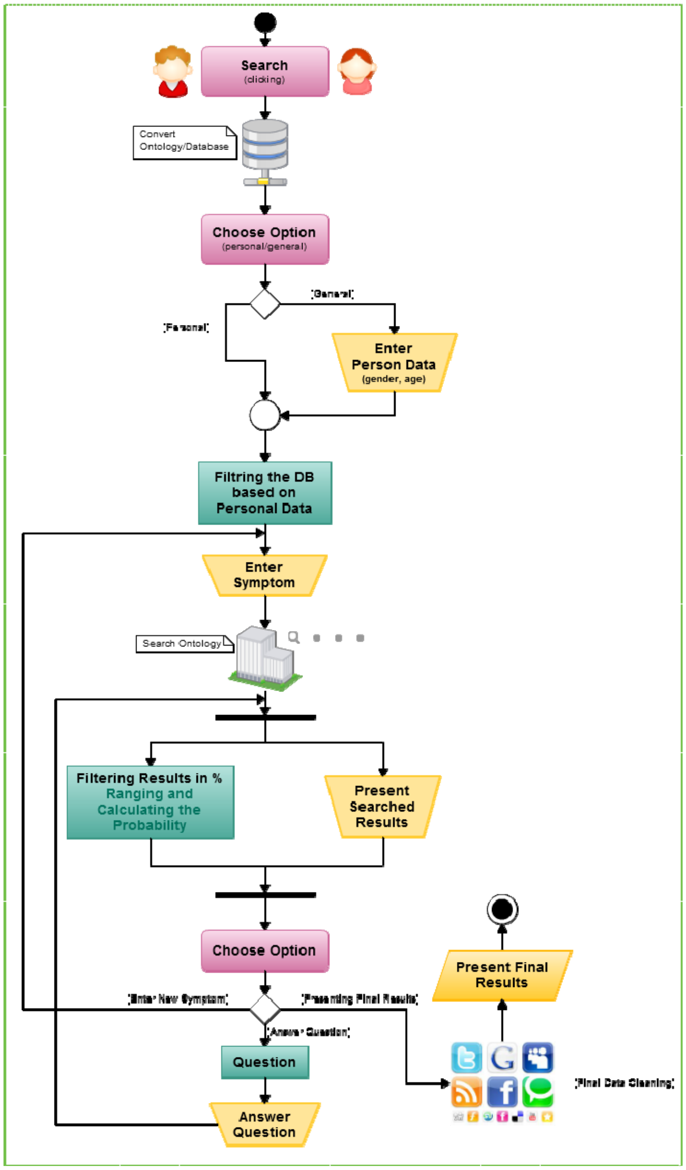

Fig. 1. MS^{2}TP Algorithm.

Social media, cell phones, and other communication modes have opened up a two-way street in health research, supplying not just a portal for delivering information to the public but also a channel by which people reveal their concerns.

Ideally, researchers want as much individualized information as they can get to anchor social network predictive models in real-world data. The power of these models was illustrated in a 2010 study by two professors and long-time collaborators—Nicholas Christakis from Harvard University and James Fowler from the University of California, San Diego, who found that social network analyses can predict flu outbreaks earlier than traditional tracking methods [20].

Hence, similar to many other scenarios, social media platforms present a weapon for health tracking. As a portal for channeling the personal experience of billions of people, they are a true reflection of our society—the good, the bad, and everything in between.

What differentiates our work from the other models presented is the multi parameters convergence in the process of predicting diseases. We describe disease predictions for a rather simple model that we named as Multiple Sources Multiple Search Techniques Prediction (MS^{2}TP), with particular focus on children diseases. MS^{2}TP is a proposed architecture that strives to optimize the processing of maximum number of data relevant for the search, aiming towards obtaining a percentage likelihood of possible outcomes with great accuracy. The application of this technique is in searching medical data, for precise determination of a diagnosis based on entered symptoms, searching the ontology, social network posts, social network profiles, demographic data, etc. Having multiple data sources and multiple search techniques will create broader view of the problem which will produce more accurate data results.

MS^{2}TP will concentrate on general children diseases that are most common and specific for our country, determined in cooperation with a doctor - pediatrician. We choose children diseases because with that we can achieve two target levels: first the parents and then the children themselves. A healthy child means a healthy adult. Every stage of human life is important but it seems that the greatest mistakes are made with children. Health complications in childhood cause predispositions for diseases in later life. So, earlier predictions could save someone's life.

In the first stage, data from the social network created for this purpose will be gathered, special prepared ontology of children general diseases and symptoms, the interaction data with users, as well as some demographical data. The future work is planned to incorporate results from medical examinations, photographs, medical records, telemedicine devices and others.

Fig. 1 shows the flow of activities of proposed search algorithm for children disease prediction. Extracting relational database from the children disease ontology is the first step of the proposed prediction algorithm. Since there is no need to frequently update the database (on each click on the search button), it will be renewed for a longer period of time (after each update of the ontology). The aim of the ontology to database mapping model is to provide access to the contents of a database through the schema of the ontology.

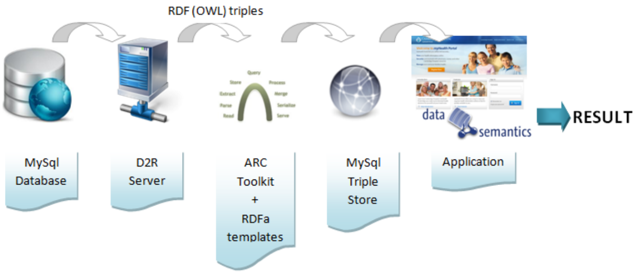

Fig. 2. Proposed architecture

From the Semantic Web point of view, the mappings are capable of corresponding class individuals (alt. instances) to any possible dataset combination, thus significantly extending the storage capability of the ontology. The proposed process of ontology-to-database conversion is divided in several phases, as shown in Fig. 2.

In the first two phases, we are generating RDF (OWL) triples from MySql data conversion using D2R Server. The resulting D2R mapping is used for mounting a SPARQL end-point that provides access to database records as RDF instances and for generating a plain ontological representation of database schemas. In the next two phases, after generation of the triples, we will use ARC toolkit (for queries) and RDFa (for visualization) in order MySql triple store to be constructed for enabling semantic data representation.

Since semantic data are in question, two endpoints are needed. The first endpoint is the machine (computer) and the second is the patient (human being). Therefore, we will have sparkle endpoint for the machine through which the data representation can be in RDF, XML, Turtle or N-triples. On the other side, for the patient as an endpoint, there will be an RDF browser or ordinary browser with embedded RDFa describable browser. The end point for the user side is needed in order for the results to be shown in readable form.

After this conversion, the process of filtering personal user data can start. In this context, two outcomes are possible: to search for ourselves or to search for some other person. The difference is that when we search for ourselves, the data from our social network profile would be used as a filter to search through the database which is not the case if we are searching for a third person. So, the user will be asked to enter some relevant personal data that will shorten the searching time and will increase the likelihood of a disease prediction. In this case, the amount of data that we would ask the user must be limited because we cannot lose his/her valuable time for entering these data.

Once we get the personal data, percentage filtering and evaluation process is performed for possible diagnoses in the database. For example, if the user is male then diseases that are common for females will be cut off from the database search by which the list of possible diagnoses will be reduced.

Data found in the user profiles would give a more realistic picture of the user and would certainly constitute an advantage over the users that do not have profile in the social network.

The next step is entering a symptom. When typing a symptom, an “auto complete” option will be offered from the ontology of particular diseases symptoms. Symptoms will be used to increase/decrease the probabilities of diagnoses depending on its relationships with the diagnosis. Once inscribed, first symptom is syntax checked. After that, synonyms for that symptom are searched and the symptom will be replaced by a characteristic symptom which is a representative or commonly used term for this symptom. The synonyms will be listed in the ontology of symptoms. For example, if the user enters *temperaura*, it would be predicted that it is *temperatura*, the Latin version of the word temperature in Macedonian language and the symptom may be appropriately presented as *temperature*, the English translation that is the most typical representative of the symptom.

Once the symptom is determined, the ontology of diseases is searched to get the relationship of that symptom with some disease and to calculate the impact of the existence of such symptom for that particular disease.

Once the ontology is searched, as a result obtained, we have sorted the database with intention to get disease possibility in terms of entered symptom. Correspondingly, the top 10 diagnoses would be shown.

After displaying the results, the user is offered a choice: new symptom to be written, questions to be asked, or final results to be shown.

- *New symptom* - in this case it is acted like the first entered symptom, taking into consideration the previous results, and the impact of symptoms as another possibility of the possible diagnoses.
- *Question* - asking questions regarding genetic predispositions, how long you have had those symptoms, typical questions for the "top 10" possible diagnoses.
- *Final results* - all possible resources are searched (posts, profile, medical history, demographic data) to obtain more reliable prediction and the final results are displayed.

Because the whole process can take a very long time, we suggest offering an option to the user to choose whether to apply time or accuracy search. The difference is in the execution time (fast/slow) and accuracy in the predicted diseases (high/low).

The ontology is specially designed for this purpose and the user will start to interact with it after he or she chooses to search for disease prediction. The ontology update will be done by the administrator from time to time, by adding new data for the diseases. So, in the search process, at the beginning, this ontology will be converted into database and after that the next steps will be rising till the end results are shown.

One can conclude that the input data will be the entered symptoms, the answers to the questions and information gathered from the social network. On the other side, the outputs will be the possible diseases presented with percentage based on the inputs from the user.

IV. Conclusion and Future Work

The WEB 2.0 technologies have been progressing at a rapid pace. They are now being called upon to support knowledge management, and not just to process data or information.

The goal of our work is to face the future of health care deliverance by proposing a system that can assist an ordinary person in decision making. For that purpose a simple model, called MS^{2}TP that will predict children general diseases with high percentage precision, is proposed.

The combination of different technologies is the results of having multiple data recourses as ontology data, social network data and others. It is at this point where the name of the algorithm has been created, which is Multiple Sources Multiple Search Techniques Prediction.

The ontology is especially designed in that way to be understandable for ordinary people and without covering complicated medical terms into its realization.

Next, we have to work on its realization and implementation and to compare our model with other proposals in the literature. The future work will combine the other medical data as patient medical records or blood results into its implementation.

References

[1] Eysenbach G. “What is e-health?”, J Med Internet Res 2001; 3: e20- doi: 10.2196/jmir.3.2.e20 pmid: 11720962.

[2] B.W. Hesse and B. Shneiderman, “eHealth research from the user’s perspective.”, Am J Prev Med, 32(5 Suppl):S97-103. [PMID:17466825], 2007.

[3] C. Courage and K. Baxter, “Understanding Your Users: A practical guide to user requirements: Methods, Tools, & Techniques”, pp. 415-456, San Francisco, CA: Morgan Kauffman Publishers, 2005.

[4] CISCO Report, “Older people, technology and community”, Accessed 11 July 2013, http://www.cisco.com/web/about/ac79/docs/wp/ps/Report.pdf.

[5] AT&T Network Report, “eHealth Initiatives Improve Patient Care, Cut Costs”, Accessed 12 July 2013, https://www.corp.att.com/stateandlocal/docs/ehealth_initiative.pdf.

[6] S.O. Marberry, “Trendspotting: The Next 10 Years of Healthcare Design”, Healthcare Design Magazine, Sep 19, 2012, http://www.healthcaredesignmagazine.com/article/trendspotting-next-10-years-healthcare-design?page=show, Accessed 13 July 2013.

[7] W3C standards repository; Semantic Web, http://www.w3.org/standards/semanticweb/; Accessed 11 July 2013.

[8] J. Mori, Y.Matsuo, K. Hashida and M. Ishizuka, “Web Mining Approach for a User-centered Semantic Web”, In Proc. Int’l Workshop on User Aspects on the Semantci Web in 2nd European Semantic Web Conf. (ESWC 2005), Heraklion, Greek, pp. 177-187, 2005.

[9] A. Stavrianou, J. Velcin and J.H. Chauchat, “A combination of opinion mining and social network techniques for discussion analysis”. Revue des Nouvelles Technologies de l'Information (RNTI), pp. 25-44, Cepadues 2009.

[10] C. Roosevelt and Jr. Mosley, "Social Media Analytics: Data Mining Applied to Insurance Twitter Posts," Casualty Actuarial Society E-Forum, winter 2012-Volume.

[11] S. Asur and B. A. Huberman, “Predicting the Future With Social Media”, roceedings of the ACM international conference on web intelligence, Toronto, 31 August-3 September 2010, pp.492-499.

[12] A. Tumasjan, T.O. Sprenger, P.G. Sandner and I.M. Welpe, “Predicting elections with twitter: What 140 characters reveal about political sentiment”, Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media, pp 178—185, 2010.

[13] V. Lampos, D. B. Tijl and C. Nello, “Flu detector - Tracking Epidemics on Twitter”, ECML PKDD 2010, Springer, pp. 599 – 602, 2010.

[14] A.V. Kshirsagar, H. Bang, A.S. Bomback, S. Vupputuri, D.A. Shoham, L.M. Kern, P.J. Klemmer, M. Mazumdar and P.A. August, “A simple algorithm to predict incident kidney disease”, Arch Intern Med. 8 December 2008, pp.168(22):2466-73.

[15] A. Sadilek, H. Kautz, V. SilenzioJacobs and C.P. Bean, “Predicting disease transmission from geo-tagged micro-blog data”, in Twenty-Sixth AAAI Conference on Artificial Intelligence, pp. 271-350, 2012.

[16] C.S. Kutur, K.R. Kanth, K. S. Kanth, “Improved Algorithm for Prediction of Heart Disease using Case based Reasoning Technique On Non-Binary Datasets”, In International Journal of Research in Computer and Communication Technology, Vol 1, No 7, December 2012.

[17] D. A. Davis, N. V. Chawla, and N. Bloom. “Predicting Individual Disease Risk Based on Medical History” Proceeding CIKM '08 Proceedings of the 17th ACM conference on Information and knowledge management, pp. 769-778, ACM New York, NY, USA, 2008.

[18] K. Steinhaeuser and N. V. Chawla, “A Network-Based Approach to Understanding and Predicting Diseases. Social Computing and Behavioral Modeling”, Springer, 209-216, 2009.

[19] R. Mooney A. Strehk and J. Ghosh, “Impact of similarity measures on web-page clustering”, In Proc of AAAI workshop on AI for Web Search, pp. 58–64, 2000.

[20] Christakis N.A. and Fowler J.H., “Social network sensors for early detection of contagious outbreaks.” PLoS One. 2010 Sep 15, 5(9):e12948.