USING SOCIAL MEDIA TO PREDICT CHILDREN DISEASE OCCURRENCE

Conference Paper · February 2014

CITATIONS
2

READS
61

3 authors:

Betim Cico
Epoka University
125 PUBLICATIONS **363** CITATIONS
SEE PROFILE

Marika Apostolova-Trpkovska
South East European University, Tetovo, Macedonia
24 PUBLICATIONS **105** CITATIONS
SEE PROFILE

Ivan Chorbev
Ss. Cyril and Methodius University in Skopje
115 PUBLICATIONS **878** CITATIONS
SEE PROFILE

USING SOCIAL MEDIA TO PREDICT CHILDREN DISEASE OCCURRENCE

Marika Apostolova Trpkovska[1], Betim Cico[1] and Ivan Chorbev[2]
[1]*South East European University - Ilindenska n. 335, 1200 Tetovo, Macedonia*
[2]*Faculty of Computer Science and Engineering - Rugjer Boshkovikj n.16, P.O. Box 393, 1000 Skopje, Macedonia*

ABSTRACT

Delivering care in the future may focus on predicting health needs rather than waiting for disease to begin. So, the future of health care may be in "predictive health" that emphasizes as much as possible prediction and as less as possible diagnoses that are coming from these predictions. Nowadays, the researchers are mining the data provided in social networks, aiming in prediction of diverse phenomena like social, political, medical, etc. We are focused on proposing a model for predicting children general diseases. 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. This model is a part of the specially designed social network where the patient medical data are not included concerning their privacy and sensitivity. Through the social network we a giving the possibility to mothers to exchange their experiences through different social services. And through the proposed disease prediction model we a giving the possibility to mothers to gain some preliminary predicted diagnosis, if needed.

KEYWORDS

Children disease prediction algorithm; ontology/database conversion architecture; semantic web; social network.

1. INTRODUCTION

Social media today supply a remarkable amount of data about users and their shared interactions. As a result they are offering many new opportunities for research exploration to computer and social scientists, economists and others. Possibly, one of the most attractive lines of research is that of predicting future events and developments based on social media data.

People are fascinated with what will happen in the future. We even associate intelligence with an ability to predict future events, according to Hawkins (2004). In the past, a number of techniques were invented including observing flight of birds, astrology, natural phenomena and others. William in the article "To err is human" claims that predictions were done mostly through experts who had developed their own intuitions and methods of prediction. Unfortunately, such expert knowledge is individual and cannot be computerized or even duplicated.

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.

Integrating social media into health activities allows peoples to leverage social dynamics and networks to encourage participation, conversation and community – all of which can help spread key messages and influence health decision making. Over the years, the Internet has changes people's relationships. Although doctors, nurses and other health professionals continue to be the first choice for most people with health concerns, online sources, including advice from peers, are a significant source of health information.

1.1 Electronic Health

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.

Hesse and Shneiderman (2007) stated that e-Health 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.

e-Health 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 (Marberry, 2012). 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.

Let's suppose that it's winter, flu season, and we are sitting in front of our computers, feeling a bit dizzy, with an unwanted swelling in the back of our throat and a headache coming on. If we are like millions of other people, we might engage in a moment of Internet enabled self-diagnosis. We pop our symptoms into a search engine, and in few seconds dozens of health-related websites appear on our screen. That search supplied us with information—some useful and some not—but in today's hyper-connected world, it also supplied a data point for those who survey disease outbreaks by monitoring how people report symptoms via social media. In fact, 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, locations, and physical movements from one place to another.

Related to the mentioned before, 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.

1.2 The Semantic Web Paradigm in Supporting Health Care

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 (Mori et al., 2005).

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 sorts 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, according to Stavrianou, Velcin and Chauchat (2009). For example, peoples' feelings about certain health decisions or reactions to particular experiences.

1.3 Social Media as a Prediction Tool

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 by Roosevelt and Mosley (2012), 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.

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. The research society has developed a lot of sophisticated techniques that aim to predict future outcomes using data-based models. Such model-based forecasts have proved to be quite successful in predicting a diversity of outcomes including economic (Clements and Hendry, 2011), societal (Silver, 2012), and political outcomes (Campbell, 2008). Despite their general success, even these models cannot predict the future perfectly, because real-world outcomes can change in ways that are not anticipated by data-based models.

The advent of social media provides researchers with a new and rich source of easily accessible data about individuals, society and potentially, the world in general. In particular, data from social media captures online behavior of users who communicate or interact on a diversity of issues and topics. It is the intent of this special section to focus on novel methods of prediction that are based on data harvested from social media. In recent years, such data has shown to be very popular with scholars interested in developing predictive models.

With varying success, an emerging community of researchers has utilized social media data for a wide variety of purposes, for example, to predict stock market movements (Bollen, Mao, and Zeng, 2011), to predict announcements of flu outbreaks (Lampos and Cristianini, 2010), to forecast box-office revenues for movies (Asur and Huberman, 2010) and even to predict election outcomes (Daniel, 2012). The models and areas of application are diverse and, moreover, predictions based on social media data have also attracted considerable attention from the public through traditional and online media. These media are projecting an impression of social media as a widely accepted and reliable source of data for predicting future outcomes.

All above mentioned has some relationship with our proposed model. We are proposing children disease prediction model based on social networking data. What differentiate our work from others is that the proposed model concentrates on percentage prediction of children general diseases based on entered symptoms, ontology search, social network posts, social network profiles, demographic data and more.

2. THE MODEL

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.

2.1 Existing Medical Related Prediction Models

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. (2008) 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. Further on, Adam Sadilek at el. (2012) 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. Next, Chandra Sheka's work (2012) deals with an improved algorithm for prediction of heart diseases using case based reasoning technique on non-binary datasets. Darcy A. Davis at el. (2008) predict individual disease risk based on medical history using a collaborative assessment and recommendation engine. Steinhaeuser and Chawla (2009) use a hybrid technique based on collaborative filtering and nearest neighbor classification. According to Silver (2012) the similarity between two patients is computed with the Jaccard coefficient, 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.

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 (Christakis and Fowler, 2010).

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.

2.2 MS^{2}TP Model

Our goal is through the proposed model, the user to have a starting idea or point of what their children could have based on the appeared symptoms (similarly to the scenario mentioned on page 2). Afterwards parents must consult a medical specialist to determine the actual disease and to continue with the treatment. Knowing that the peoples' (patients') medical data are very crucial, concerning their privacy and security, they are not included in our model.

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. MS2TP 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.

MS2TP 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.

Social networking has evolved from chatting on Mirc and Yahoo Messenger to niche networking sites for people with all kinds of interests. Lately MySpace, Facebook and even more recently Twitter have exploded

on the scene and social networking is as common as talking on the phone. In fact, the combination of text messaging and online social networking has certainly given the phone a run for its money.

Social networking sites like Facebook cater to a smorgasbord of users while people using sites such as Subeta and Neopets are mostly teens. Social networking sites can be very specific, too. For instance, you'll find sites for just about every type of car, sport and sports team or music genre. And one of the largest groups of specialized social media sites is dedicated to motherhood and parenting.

Figure 1. Social network features

Therefore, we are going to create a social network specialized for mothers. This network as shown in Fig. 1, will enable mothers to stay in touch, share parenting advice, review products, create groups, discuss topics we all face as a parent, real time chatting, activate different health oriented services like vaccination calendar and others, plus it will have implemented an specialized Children Disease Prediction Engine based on the proposed MS2TP model. This model makes our network different from others.

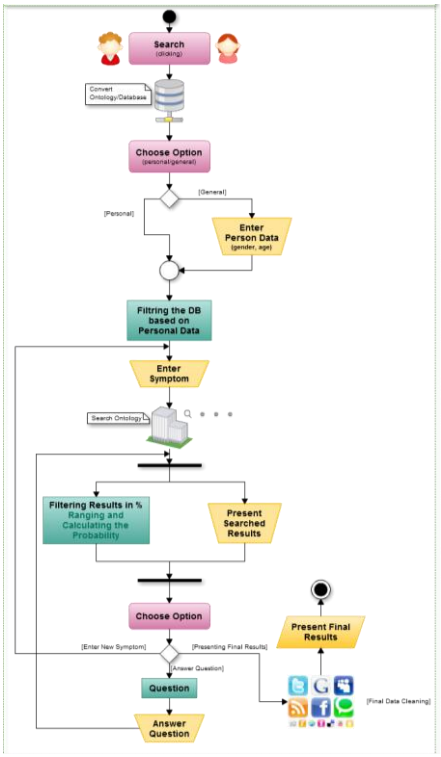

Figure 2. MS2TP algorithm

We know that motherhood can be a hard job especially when mums have to juggle work and family so we will create a lot of interactive features enabling mums to share practical information.

In the first stage, data will be gathered from:

- the social network created for this purpose,
- the special prepared ontology of children general diseases and symptoms,
- the interaction questions/answers with users,
- the users' demographical data.

Fig. 2 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.

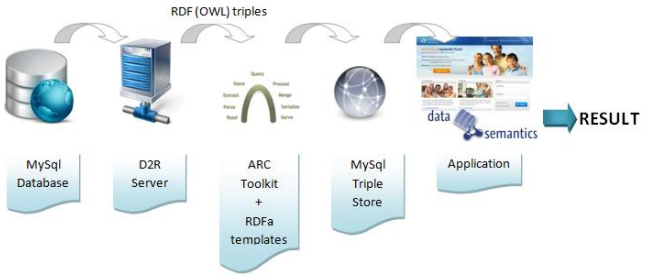

Figure 3. Proposed architecture

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. 3.

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. The ontology is shown on Fig.4.

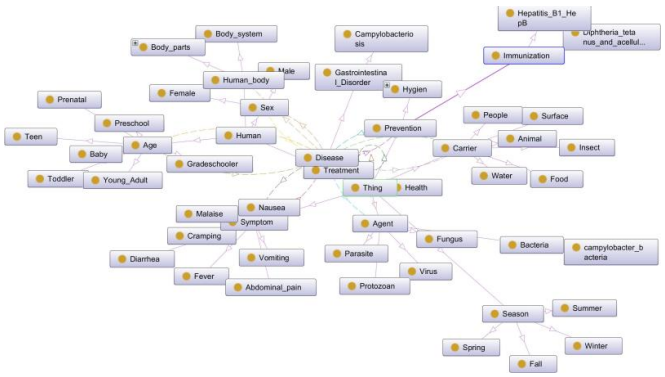

Figure 4. Ontology

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.

3. 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 social media and other communication modes increase the people knowledge about health issues. In our case the proposed model empowers people to make safe health related pre-decisions. We must stress that our proposal do not replace the medical specialist judgment. It is just a middle pre-decision point between the people and the medical specialist.

So, 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 MS2TP 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.

As a future work we have to work on model realization and implementation and to compare our model with other proposals in the literature.

REFERENCES

A.V. Kshirsagar. et al, 2008. A simple algorithm to predict incident kidney disease. *Arch Intern Med*. pp.168(22):2466-73.

A. Sadilek. et al, 2012. Predicting disease transmission from geo-tagged micro-blog data. *Twenty-Sixth AAAI Conference on Artificial Intelligence*. Toronto, Ontario, Canada, pp. 271-350.

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

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

Bollen, J. et al, 2011. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1), pp. 1-8.

Campbell, J.E., 2008. The trial-heat forecast of the 2008 presidential vote: performance and value considerations in an open-seat election. *PS: Political Science and Politics*. Vol. 41 No. 4, pp. 697-701.

Christakis N.A. and Fowler J.H., 2010. Social network sensors for early detection of contagious outbreaks. *PLoS One.* 5(9):e12948.

Clements, M.P. and Hendry, D.F., 2011. *The Oxford Handbook of Economic Forecasting*, Oxford University Press, Oxford.

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

C.S. Kutur, K.R. et al, 2012. Improved Algorithm for Prediction of Heart Disease using Case based Reasoning Technique On Non-Binary Datasets. *International Journal of Research in Computer and Communication Technology*, Vol 1, No 7.

D. A. Davis. et al, 2008. 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.

Gayo-Avello, Daniel. 2012, A meta-analysis of state-of-the-art electoral prediction from Twitter data, arXiv:1206.5851 [cs.SI]

Hawkins, J., 2004. *On Intelligence*. Times books. ISBN 0805074562.

J. Mori, et al, 2005. Web Mining Approach for a User-centered Semantic Web. *Proc. Int'l Workshop on User Aspects on the Semantci Web in 2nd European Semantic Web Conf. (ESWC)*, Heraklion, Greek, pp. 177-187.

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

Lampos, V., and Cristianini, N., 2010. Tracking the flu pandemic by monitoring the Social Web. *Proceedings of the 2nd IAPR Workshop on Cognitive Information Processing (CIP 2010)*, pp. 411-416, IEEE Press.

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

S.O. Marberry, 2012. Trendspotting: The Next 10 Years of Healthcare Design. *Healthcare Design Magazine*. http://www.healthcaredesignmagazine.com/article/trendspotting-next-10-years-healthcare-design?page=show.

Silver, N., 2012. The Signal and the Noise. The Art and Science of Prediction. *Allen Lane*, London.

S. Asur and B. A. Huberman, 2010. Predicting the Future With Social Media. *Proceedings of the ACM international conference on web intelligence*, Toronto, pp.492-499.

William J. *To Err is Human*. http://www.williamjames.com/Science/ERR.htm.