Social Health Signals

Ashutosh Jadhav*, Swapnil Soni*, Amit P. Sheth

[ashutosh, swapnil, amit]@knoesis.org

Kno.e.sis center, Wright State University, Dayton OH, USA
* First Author

Abstract. Recently Twitter, has emerged as one of the primary medium for sharing and seeking of the latest information related to variety of the topics including health information. Although Twitter is an excellent information source, identification of useful information from the deluge of tweets is one of the major challenge. Twitter search is limited to keyword based techniques to retrieve information for a given query and sometimes the results do not contain real-time information. Moreover, Twitter does not utilize semantics to retrieve results. To address these challenges, we developed a system (Social Health Signals) by leveraging rich domain knowledge to extract relevant and reliable health information from Twitter in near real-time. We have used semantics based techniques 1) to retrieve relevant and reliable health information shared on Twitter in real-time 2) to enable question answering 3) to rank results based on relevancy, popularity and reliability 4) to semantically categorize the information.

Keywords: Twitter, Semantic Web, Social Media Analysis, Text Mining, Health Informatics, Real-time Data Processing

1 Introduction

Over the past decade, Twitter, has become a primary mode for users to share and seek information on different topics, including health information. According to a consumer survey, one-third of the consumers now use social media for seeking medical, tracking and sharing health information [1]. Twitter allows users to create 140 character messages (tweets) with an option to include weblink to share health information publicly. This health information can be useful and educative resource for others. On the Twitter, more than 75,000 worldwide healthcare professionals post 152,000 tweets every day[2]. In some cases people prefer Twitter as a information source compared to traditional information sources since they can find all the timely information aggregated at one place. In our study, we have used Twitter as a data source and one of the most common chronic diseases, diabetes, as a use case.

Online Health Information Seeker (OHIS) have different preferences when it comes to find out information related to health conditions through social media search [3]. Some OHIS prefer real-time (latest) information, breaking news

(articles), while others prefer facts and the information that contributes to general understanding of a health condition [4] [3], etc. Consider a scenario where a diabetic patient, John, is interested in keeping himself up-to-date with latest information about diabetics. How can he do this? Here John can leverage the strengths of Twitter platform on which almost all the important health information related to diseases, drugs, clinical trials, side effects are being shared. Twitter has provided a search option but it posses following significant challenges: keyword based techniques are used for search result retrieval, semantics of the query is not considered, sometime results do not contain real-time information, and ranking the results does not considered reliability and popularity factors.

To address the limitations of Twitter search and to overcome Twitter's information overload challenge, we have build a system, (Social Health Signals - SHS), where 1) reliable and popular health information from Twitter for a topic is aggregated 2) users can ask health related questions 3) to enable efficient browsing of the results, by semantic health categories such symptom, food and diet, healthy living and prevention 4) location and volume based visualization of the tweets 5) to complement dynamic health information from Twitter SHS also provides static (factual) information about disease from Wikipedia. The techniques used in the implementation of this system are principally based on domain semantics, knowledgebases (UMLS, WordNet) and Semantic Web techniques. For example, we used taxonomy based approach for a) data collection b) search query understanding c) data annotation and retrieval. We have also used ontological knowledge and domain knowledge from UMLS to perform semantic categorization of health information into health categories.

2 System Architecture

The system is divided into four major components: data processing pipeline, pattern extractor, rank calculator and semantic categorizer.

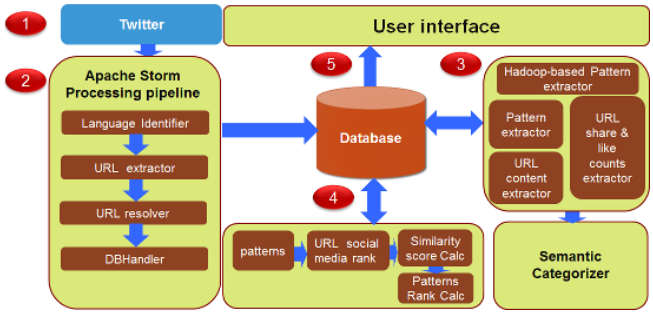

Fig. 1. Architecture diagram of Social Health Signals platform

2.1 Data Processing Pipeline

To extract features from real-time tweets, the first challenge is to create a infrastructure to collect real-time tweets. In our system, we have create a processing pipeline using Apache Storm to collect the real-time tweets and associated metadata using the public Twitter streaming API. We have used the Twitter streaming API to crawl real-time tweets. To collect the data related to a disease, we have used taxonomy based approach in which we collected terms associated with disease from UMLS, domain experts and by doing empirical study. To extract theses features and other metadata such as mapping location information, short url, etc. from the tweets in real-time, we have used different Apache storm's components to implement the logic.

2.2 Question and Answering on Twitter data

One of the feature of this a system to let users to ask health-related questions on Twitter data. Here, we have divided users questions into two categories: static and dynamic. The static questions are preselected frequently asked questions collected from the different sources such as Mayo Clinic, WebMD, etc. The dynamic questions are typed by the user on the fly. To extract relevant document, we have used triple based pattern (subject, predicate, and object) mining technique to extracts triple patterns from tweets. The triple pattern is defined in the initial question. We have used an AQL tool to construct triple-patterns, and Apache Hadoop Map-Reduce framework for speedup. To expand the query (or triple), we have incorporated the domain knowledge using UMLS-Metathesaurus (Unified Medical Language System) and WordNet.

2.3 Ranking

We have used the following features to rank the results are: popularity, relevancy, and reliability. To check the popularity of URLs through social media (e.g., a Twitter and a Facebook) share and like counts. Similarly, for reliability we use the URLs Google domain pagerank. In our approach, we have used a TF-IDF cosine similarity algorithm. Once all the features are extracted, we have evaluated many machine learning algorithms and selected "Random Forest algorithm" based on an evaluation matrix (Normalized discounted cumulative gain).

3 Semantic Categorisation

To enable efficient browsing of the health information, we categories tweets and new articles into health categories. We have used a rule-based categorization approach developed by Jadhav et al. [6] [7]. First the tweets and new articles are annotated with UMLS concepts and semantic types using UMLS MetaMap. Each health category has certain UMLS concepts and semantic types which are used as a rule for the categorization. After health categorization, users can browse the information based on the health categories

4 User Interface

Social Health Signals (demo link) has four widgets are: explorer, top 10 articles extractor, tweet traffic explorer and tag cloud. The explorer widget allow users to select a most frequently question or type a new question (dynamic) for extracting information (semantic categorised) in near real-time. Similarly, top articles extractor shows the popular news articles which has been much shared and liked by the Twitter's users. A tweet traffic explorer widget shows (heatmap) locations of tweets collected every six hours. Finally, tag cloud widget is useful for users to know the popular keywords.

5 Evaluation, Discussion and Conclusion

We have selected reliability, relevancy, and real-time features for the evaluation of SHS results with Twitter search. We conducted three qualitative focus group studies to access performance of SHS with respect to Twitter search. In all three studies, user preferred content from SHS over Twitter search. To find useful health information in real-time from Twitter, there are many challenges such as the real-time nature of Twitter, information overload and noisy data. We have dealt with each of the challenges by using state-of-the-arts technologies and semantic Web based techniques in our system. This a very comprehensive system with motivation to aid users to keep track of health information. The system developed with contribution to public health systems as well as social media based systems.

References

1. Ottenhoff: Infographic, Rising Use of Social and Mobile in Healthcare, The Spark Report (2012)
2. Ilene MacDonald: Healthcare professionals flock to Twitter,FierceHealthcare
3. Choudhury et al.:Seeking and sharing health information online: Comparing search engines and social media, ACM (2014)
4. Teevan et al.: # TwitterSearch: a comparison of microblog search and web search, The Spark Report. ACM 2011
5. IBM: InfoSphere Streams text analytics, "`http://www.ibm.com/developerworks/library/bd-streamstextanalytics/`", 2015
6. Jadhav et al. Analysis of Online Information Searching for Cardiovascular Diseases on a Consumer Health Information Portal, AMIA Annual Symposium 2014
7. Jadhav et al. Online Information Seeking for Cardiovascular Diseases: A Case Study from Mayo Clinic at 25th European Medical Informatics Conference, 2014
8. Sheth et al. Twitris- a System for Collective Social Intelligence. Encyclopedia of Social Network Analysis and Mining (ESNAM), 2014.
9. Jadhav et al. Twitris 2.0: Semantically Empowered System for Understanding Perceptions From Social Data , Semantic Web Application Challenge at ISWC, 2010