Domain Specific Document Retrieval Framework for Real-time Social Health Data

Swapnil Soni
swapnil@knoesis.org

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

Abstract. With the advent of the web search and microblogging, the percentage of Online Health Information Seekers (OHIS) using these online services to share and seek health real-time information has increased exponentially. OHIS use web search engines or microblogging search services to seek out latest, relevant as well as reliable health information. When OHIS turn to microblogging search services to search real-time content, trends and breaking news, etc. the search results are not promising. Two major challenges exist in the current microblogging search engines are keyword based techniques and results do not contain real-time information. To address these challenges, we developed an approach to search near real-time and reliable content from Twitter, based on triple-pattern mining, near real-time retrieval, and ranking considering popularity and relevancy of the results.

Keywords: Twitter, Data mining, Triple pattern, Real-time, Health, Chronic disease, Social Media analysis, Text mining

1 Introduction

Over the past ten years, percentage of social media users has increased exponentially. In the U.S, 72% of online users use social media and its popularity grown by 64% since 2005 [1]. Social media has become primary mode for users to share and find 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 [2]. A popular service, Twitter, allows users to create tweets and optionally include links in the tweets to share health information publicly. This health information can be useful for others to learn from the shared information. On the Twitter, more than 75K worldwide healthcare professionals post 152K tweets every day[6]. In our study, we have used Twitter as a data source and one of the most common chronic diseases, diabetes, as a use case.

1.1 Background and Motivation

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. OHIS have many options on the Internet for health information seeking in real-time such as Google time-bound search, Twitter search, etc. But search results from these venues possess some significant challenges: the results are not real-time, search results are based on keyword-based techniques, and ranking of the results based on a relevance to each individual keyword in the query. A leading microblog search service such as Twitter use keyword-based approach, and since the Twitter is overloaded with information, and merely matching query keywords with tweets to locate relevant set of documents of information is inappropriate. Also, we observed that in Twitter search the results are not near real-time due to the keyword-based relevancy algorithm. Furthermore, Twitter search does not use domain knowledge and reliability factors to rank the results.

The objective of this research is to build a system for users to ask health-related questions to obtain reliable, and relevant health information shared on social media in near real-time. But, how to extract near real-time, reliable and relevant documents from the health information shared on a Twitter for a given user query? To extract relevant documents from a Twitter in near real-time based on a given user query, we have to deal with real-time tweets, information overload, and noisy data.

2 Related Work

2.1 Microblog Retrieval Method

The amount of conversation on Twitter has increased exponentially over last decade. To address the Twitter's information overload challenge, many researchers use microblogging services like Twitter to find out health information;however, extracting useful information is challenging given its volume, inconsistent writing, and noise. To extract useful information from a Twitter, many researchers worked on various retrieval models such as a user-based tree model, term-based, and pattern-based approaches. A Twitter based social media analytics system, Twitris uses Spatio-Temporal-Thematic (STT) processing of the Twitter data [9] [10]. However, many researches favor term-based extraction model also known as keyword based extraction. The keyword based model extract information based on keyword matching of users query. Its possible to extract undesired results based on a user query due to keyword based model extraction.

Magnani et al. proposed a term-based model for retrieving conversations from microblogs [5]. In this study, the concept of conversation retrieval from Twitter, a preliminary version of the concept presented by Magnani et al., proposes a user-based tree model to retrieve conversations from microblogs [5]. In this research, the whole conversation of users are represented as a tree, and its message and reply are represented as nodes. These conversations are stored in IR engine Lucene for indexing the text which can help the system to retrieve the relevant conversation documents based on the query. After finding the rel-

evant conversation, the system ranks the relevant conversations based on text relevance, popularity, timeliness, audience, and density features.

3 Data Collection and Feature Extraction

In this study, we have used tweets (messages shared on Twitter) and URLs content (for URL(s) mentioned in the tweets) as the data sources to extract relevant information for a given user query. To extract features from real-time tweets, the first challenge is to create a infrastructure to collect real-time tweets. In our research, we have used Apache Storm to collect the real-time tweets using the public Twitter streaming API while also performing meta-data extraction. Apache storm is, open source software, used for real-time, distributed computing. Spouts and Bolts are the basic components in the storm for real-time processing of the data. The bolts contain computation logic to perform features extraction logic in real-time.

A tweet has many features, such as text, short url, latitude or longitude, retweet count, etc. All these features are useful for finding out useful information. To extract theses features from the tweets in real-time, we have used bolts (a Apache storm's components) to implement the logic. This process is also known as a pre-processing pipeline in our system.

4 Extraction of Relevant Documents

The objective of this research is to build a system to ask health-related questions on Twitter data. Hence, we have divided users questions into two categories: static and dynamic. The static questions are preselected frequently asked questions collected from the different sources. Also, the dynamic questions are typed by the user on the fly, which is not the case with the static queries. We proposed a novel approach by extracting real-time tweets, pattern-mining, incorporating domain knowledge, and including popularity measures of the content (tweets + URLs) in ranking of the results.

To make the system near real-time, the search results are divided into intervals of six hours. The near real-time process of extracting relevant documents is depends on static and dynamic questions are different. In the case of static questions, we extract documents every six hours, while in dynamic questions, we extract documents from that moment to last six hours data. To extract document, we have used triple based pattern (subject, predicate, and object) mining technique to extracts triple patterns from microblog messages–related with chronic health conditions. The triple pattern is defined in the initial question. To extract information or documents we have used the IBM text analytic tool AQL (Annotated Query Language). AQL is a query language to help developers to build queries that extract structured information from unstructured or semi-structured text. We have used an AQL tool to construct triple-patterns, and for faster processing we implemented it on Apache Hadoop Map-Reduce framework. To expand the query (or triple), we have Incorporated the domain

knowledge using UMLS-Metathesaurus (Unified Medical Language System) and WordNet. UMLS is used to collect authentic and reliable vocabularies related to health and biomedical. Similarly, we used the WordNet to get the synonyms of the tokens (non medical term). Furthermore, in addition to tweets, we use URLs (mentioned in the tweet) content as the data source.

5 Ranking

To simply receiving results, users want the results to be good quality, reliable and well-ordered. Existing microblog search engines (e.g., Twitter) focused on ranking algorithms to order the results based on relevance to each individual keyword in the query. 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 (filtration criteria is pagerank greater than 4). Also, we have used the relevance of the documents based on the similarity score. 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 one of them based on an evaluation matrix (Normalized discounted cumulative gain). The algorithm we have chosen is the "Random Forest" algorithm.

6 Evaluation

As our research is focused on extracting near real-time health information based on users search queries, we have made the decision to evaluate our systems results with existing real-time search engine is a Twitter. We have selected reliability, relevancy, and real-time factors to measure our results with Twitter. To evaluate the reliable source, we compared a Google domain pagerank of our top 10 results with the Twitter's top 10 results. Also, for real-time we have compared the Twitter search results with our system's search results. We found that Twitter search results are not real-time as compared to our results (which is six hours of data). Similarly, we conducted three surveys to check the relevance of the results in which we selected three questions dealing with the chronic disease diabetes. The questions are "How to control diabetes?", "What are the causes of diabetes?", "What are the symptoms of diabetes?". Upon completion of the the surveys, for all the queries 50%, 60%, and 50% of users ranked the quality of our results as "very good", whereas the results were 40%, 10%, and 40% for a Twitter search results.

7 Discussion and Conclusion

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 a novel approach in our system. Also, we have used URL's content for finding information because the tweets contains less information. However, the system does not extract factual answers of a user questions. In this thesis, I am extracting relevant documents based on a user query in near real-time. We want to extend this thesis further by including semantic categorisation in which the results will be categorised (drug, medication, symptom , etc) using prior work [7][8].

Twitter has changed the traditional way of sharing and seeking health information by health-care professionals and the general users. All kinds of information are available on the Internet for each type of user. We have tried to resolve the challenges for those who want the latest information. Our system provides a platform to users to use Twitter for finding relevant documents based on a user's question in near real-time.

8 Acknowledgement

I would also like to thank my thesis advisor Prof. Amit Sheth from Kno.e.sis center, Wright State University and my mentor Mr. Ashutosh Jadhav for their support and guidance.

References

1. Brenner and Smith: Online Adults are Social Networking Site User, Pew Research Center Internet (2013)
2. Ottenhoff: Infographic, Rising Use of Social and Mobile in Healthcare, The Spark Report (2012)
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. Magnani et al.:Advances in Information Retrieval, Springer (2011)
6. Ilene MacDonald: Healthcare professionals flock to Twitter,FierceHealthcare
7. Ashutosh Jadhav, Amit Sheth, Jyotishman Pathak Analysis of Online Information Searching for Cardiovascular Diseases on a Consumer Health Information Portal, AMIA Annual Symposium 2014
8. Ashutosh Jadhav, Stephen Wu, Amit Sheth, Jyotishman Pathak Online Information Seeking for Cardiovascular Diseases: A Case Study from Mayo Clinic at 25th European Medical Informatics Conference, 2014
9. A. Sheth, A. Jadhav, P. Kapanipathi, C. Lu, H. Purohit, G. A. Smith, W. Wang. Twitris- a System for Collective Social Intelligence. Encyclopedia of Social Network Analysis and Mining (ESNAM), 2014.
10. A. Jadhav, H. Purohit, P. Kapanipathi, P. Ananthram, A. Ranabahu, V. Nguyen, P. Mendes, A. G. Smith, M. Cooney, A. Sheth (2010), Twitris 2.0: Semantically Empowered System for Understanding Perceptions From Social Data , Semantic Web Application Challenge at ISWC, 2010