Domain Specific Document Retrieval Framework on Near Real-time Social Health Data

A thesis submitted in partial fulfillment
of the requirements for the degree of
Master of Science

by

Swapnil Soni
B.E., Jabalpur Engineering College, 2008

2015
Wright State University

Wright State University
GRADUATE SCHOOL

September 5, 2015

I HEREBY RECOMMEND THAT THE THESIS PREPARED UNDER MY SUPERVISION BY Swapnil Soni ENTITLED Domain Specific Document Retrieval Framework on Near Real-time Social Health Data BE ACCEPTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF Master of Science.

Dr. Amit P. Sheth
Thesis Director

Mateen M. Rizki, Ph.D.
Chair, Department of Computer Science and Engineering

Committee on
Final Examination

Dr. Amit P. Sheth

Dr. Krishnaprasad Thirunarayan

Dr. Tanvi Banerjee

Robert E.W. Fyffe, Ph.D.
Vice President for Research and
Dean of the Graduate School

ABSTRACT

Soni, Swapnil. M.S., Department of Computer Science and Engineering, Wright State University, 2015. *Domain Specific Document Retrieval Framework on Near Real-time Social Health Data.*

With the advent of web search and microblogging, the percentage of Online Health Information Seekers (OHIS) using these services to share and seek health information in real-time has increased exponentially. Recently, Twitter has emerged as one of the primary mediums for sharing and seeking of the latest information related to a variety of topics, including health information. Although Twitter is an excellent information source, the identification of useful information from the deluge of tweets is one of the major challenges. Twitter search is limited to keyword-based techniques to retrieve information for a given query and sometimes the results do not contain up-to-date (real-time) information. Moreover, Twitter does not utilize semantics to retrieve results. To address these challenges, we developed a system termed 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 to

- retrieve relevant and reliable health information shared on Twitter in real-time,
- enable question answering,
- to rank results based on relevancy, popularity and reliability, and
- to enable efficient browsing of the results, we group the search results into health categories using domain knowledge (semantic categorization)

In our approach, we have considered Twitter to search documents based on several unique features, including triple-pattern based mining, near real-time retrieval, and tweet contained URL based search. First, the **triple-based pattern** (subject, predicate, and object) mining technique extracts triple patterns from microblog messages related to chronic

health conditions. The triple pattern is defined in a user given question (natural language). Second, in order to make the system near **real-time**, the search results are divided into intervals of six hours. Third, in addition to tweets, we **use the content of the URLs** (mentioned in the tweet) as the data source. Finally, the **results are ranked according to relevancy and popularity** such that at a particular time the most relevant information for the questions are displayed instead of basing results solely on temporal relevance. Our evaluation focuses on questions related to diabetes, such as "How to control diabetes?, "and compare the results with a Twitter search. To measure our results with Twitter, we have selected reliability, relevancy, and real-time features for the evaluation. We have conducted a blind survey to check the relevance of the results in which we selected three questions dealing with diabetes. 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 timestamp of the Twitter search results with our system's search results.

Contents

1 Introduction **1**
- 1.1 Background and Motivation . . . 2
- 1.2 Challenges . . . 4
- 1.3 Approach . . . 5
- 1.4 Evaluation . . . 7

2 Related Work **8**
- 2.1 Mining Twitter information . . . 8
- 2.2 Usage of Twitter in Healthcare . . . 10
- 2.3 Comparison of Microblog Search and Web Search . . . 12

3 Data Collection and Feature Extraction **15**
- 3.1 Data Source . . . 15
  - 3.1.1 Twitter . . . 15
  - 3.1.2 Twitter Streaming API . . . 17
  - 3.1.3 Unified Medical Language System (UMLS) . . . 18
- 3.2 Tweet Retrieval and Feature Extraction . . . 18
  - 3.2.1 Architecture . . . 19
  - 3.2.2 Apache Storm . . . 19
  - 3.2.3 Feature Extraction . . . 20
  - 3.2.4 Dataset . . . 23

4 Extraction of Relevant Documents **25**
- 4.1 Introduction . . . 25
- 4.2 Apache Hadoop . . . 28
- 4.3 Information Extraction . . . 28
  - 4.3.1 Extraction of URL Content . . . 29
  - 4.3.2 Social Medias URL Shares and Like Counts . . . 30
- 4.4 Extraction of Triple-Pattern . . . 30
  - 4.4.1 Triple . . . 31
  - 4.4.2 Annotation Query Language (AQL) . . . 31

5 Ranking of Results **33**
5.1 Features Selection 34
5.1.1 Popularity Features 34
5.1.2 Relevancy Features 35

6 Result and Evaluation **38**
6.1 Survey 38
6.2 Results 40
6.3 Evaluation Matrix 43
6.4 Evaluation 44

7 Discussion and Future Work **47**
7.1 Discussion 47
7.2 Future Work 48
7.3 Conclusion 49

Bibliography **51**

List of Figures

1.1 Twitter result for “How to control diabetes”query shows that results are not latest . . . 3
1.2 Google result for “How to control diabetes ”query shows that results are duplicated (different web links) . . . 4
1.3 Architecture diagram of Social Health Signals platform: It has three major components- an Apache storm pipeline, a Hadoop based pattern extractor, and Semantic categorizer. . . . 6

3.1 Tweet retrieval and feature extraction architecture . . . 19
3.2 Storm’s Bolt and Spot topology . . . 20
3.3 Feature extraction pipeline . . . 21

4.1 Architecture diagram of Social Health Signals platform: It has three major components- an Apache storm pipeline, a Hadoop based pattern extractor, and Semantic categorizer. . . . 26

5.1 Comparison of ranking algorithms based precision and recall . . . 37

6.1 A survey example for “How to control diabetes? ” . . . 39
6.2 Twitter filters search results based on keywords, and the results are not latest (e.g., 6h, 7h, 15h) . . . 40

List of Tables

1.1 Apache storm pipeline- Analytic components 6

3.1 Apache storm pipeline- Feature extraction components 22
3.2 Tweet metadata 23
3.3 URL's metadata (present in a tweet) 24

6.1 Survey results of a "high quality "only 42
6.2 Survey results of a "good quality "only 42
6.3 Survey results of a "bad quality "only 43
6.4 Evaluation matrix (nDCG) for a query 1 45
6.5 Evaluation matrix (nDCG) for a query 2 46
6.6 Evaluation matrix (nDCG) for a query 3 46

Acknowledgment

I would like to take this opportunity to extend my sincere gratitude to my advisor Prof. Amit Sheth for the continuous support of my master thesis study and related research, for his motivation, and immense knowledge. His guidance helped me in all the time of research and writing of this thesis.

Besides my advisor, I would also like to extend my sincere thank the rest of my thesis committee: Prof. T.K. Prasad, and Dr. Tanvi Banerjee for their insightful comments and encouragement. Last but not the least, I would like to thank my mentor Ashutosh Jadhav for supporting me throughout writing the thesis. I could not have imagined without him to complete it.

This material is based upon work supported by the National Science Foundation under Grant IIS-1111182 "SoCS: Collaborative Research: Social Media Enhanced Organizational Sensemaking in Emergency Response." Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Introduction

Over the past decade, the percentage of social media users has increased sharply. In the U.S, 72% of online users used social media, and its popularity has grown by 64% since 2005 [6]. Social media has become the primary medium for users to express opinions on different topics, get news, and share status updates. Also, people use social media to share about the events, ones life, and personal health information. This type of content, also known as user-generated content, is largely open to public (e.g,, Twitter posts). One of the most frequent topics of such shared information is health. According to one consumer survey, one-third of the consumers now use social media for seeking, tracking and sharing health and medical information [24], while another source indicates that more than 40% of consumers say that information found via social media affects the way they deal with their health [14]. Patients of all ages are increasingly using social sites to share, seek, and engage with others who are discussing health-related topics. A popular service like Twitter allows users to create tweets and, optionally, contribute links to share health information publicly. This health information can be useful for others to learn from the shared information.

This section will explain the background and motivation of our research. Subsequently, we will define the problem and state the objective. Finally, the research methodology will be described.

1.1 Background and Motivation

According to the Pew Research Center, 45% percent of the U.S. adults are affected by one or more chronic conditions, such as diabetes, high blood pressure, and cancer [8]. The research center also showed that 53% of US adults suffering from one or more chronic diseases share and seek health information online [13]. Similarly, 66% of adults with no chronic diseases also use the Internet to collect health information [13]. OHIS have different preferences when it comes to finding out information related to health conditions through social media search [11]. Some OHIS prefer timely health information, breaking news (articles), while others prefer facts and the information that contributes to general understanding of a health condition [29] [11], etc. On the social media, OHIS can follow health professionals to get latest health information as well as share their health experiences, interesting health information, and web articles.

Recently, Twitter has become the primary medium for OHIS to share and seek information on different topics, including health information. Twitter allows users to create 140 character messages (tweets) with an option to include a web link to share health information publicly. This health information can be useful and an educative resource for others. On Twitter, more than 75,000 healthcare professionals worldwide post 152,000 tweets every day[20]. In some cases, people prefer Twitter as an information source compared to more traditional information sources since they can find timely information aggregated in one place [14]. Consider a scenario where a diabetic patient, John, is interested in keeping himself up-to-date with the latest information about diabetics. How can he do this? Here, John can leverage the strengths of the Twitter platform on which almost all the important health information related to diseases, drugs, clinical trials, and side-effects are being shared. Twitter has provided a search option, but it poses the following significant challenges:

- keyword based techniques are used for search result retrieval,
- the ranking the results does not consider reliability and popularity factors.
- often results do not contain real-time information, and
- no considerations of the categorization of search result based on semantic– user may use non technical term so an results using technical term will be ignored,

Similarly, the leading web search engine, Google, provides filtered results, but the results are not real-time and are often repeated. To find out the repetition in Google, we have performed the search (“cause of diabetes ”) over different days and found that consistently search results are almost same. Since the Internet is overloaded with information, merely matching query keywords with web pages to locate a relevant set of documents of information is inappropriate [25]. Similarly, Googles time-bound search and Twitter search results sometimes dominated with breaking news (Figures 1.1, 1.2). For example, we searched “How to control diabetes ”in Googles time-bound and observed that the top results are dominated with the same content but with different web links.

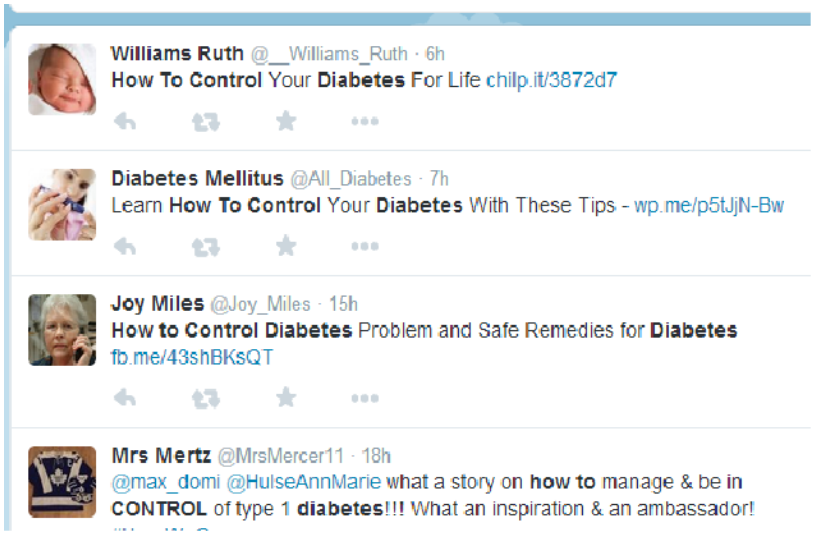


Figure 1.1: Twitter result for “How to control diabetes”query shows that results are not latest

Figure 1.2: Google result for "How to control diabetes "query shows that results are duplicated (different web links)

1.2 Challenges

Twitter data is unstructured or semi-structured, and finding out relevant documents using Twitter search is a tedious task. Many researchers have worked on retrieval of tweets based on a user query, but the results are not promising as all the popular attempts at Twitter data-mining are limited to a keyword-based analysis. Also, in other approaches, researchers use Twitter features like hashtags or any metadata information to mine tweets. However, not all posts are marked with hashtags, and people use different language to annotate tweet with hashtag [15] [7]. For example, during the swine flu or H1N1 pandemic, Chew and Eysenbach collected two millions tweets. These tweets contained sharing resources, personal stories, interest, opinion, humor, frustration, concern, relief, misinformation, and questioning, and also found that a small percentage (4.5%) was classified as possible misinformation or speculation. Furthermore, it was discovered that 90% contained web links to news or some other form of information [9]. This shows that the healthcare tweets contain a lot of information, but the challenge is to mine useful information in near real-time.

There are various retrieval models on mining information from microblogs, but the most frequently used models are keyword-based. Therefore, results are not promising. Furthermore, the real-time nature of Twitter creates a problem for the extraction of information due to information overload.

1.3 Approach

To address the limitations of Twitter search and to overcome Twitter's information overload challenge, we have built 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, knowledge-bases (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. Figure 1.3 shows the architecture schematic.

The components are described next:

1) Apache Storm Pipeline: It is used to collect real-time tweets and to perform analysis via the Twitter streaming API and Apache Storm, respectively. The Twitter streaming API uses keywords to filter tweets.

2) Pattern extractor: This module is used for extracting relevant information or documents. To extract these relevant information, we have used IBM text analytic Annotated Query

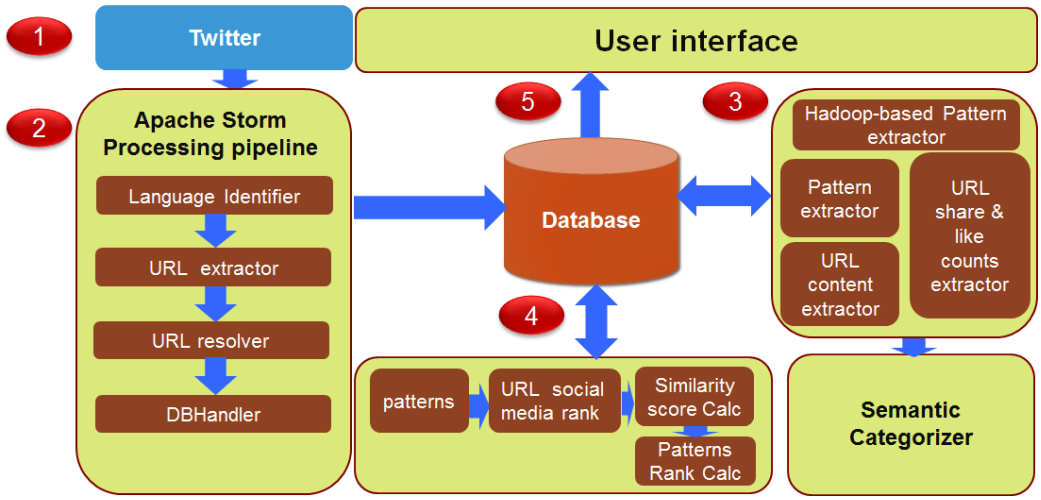


Figure 1.3: Architecture diagram of Social Health Signals platform: It has three major components- an Apache storm pipeline, a Hadoop based pattern extractor, and Semantic categorizer.

Table 1.1: Apache storm pipeline- Analytic components

Component name	**Description**
Language identification	For filtering out non-English tweets
Crawler	It crawls the real time tweets from Twitter based on the keywords
Hashtag retrieval	It is used for retrieving hashtags from the tweets
Informative analysis	Analyzing how informative is a tweet
URLs resolver	Expanding the URLs(weblinks in the tweets) from short to its original form
URLs extractor	Extract URLs(weblinks in the tweets) from tweets
Location retrieval	To retrieve geo-coordinates of the tweets
DBHandler	To save all the extracted features and tweet object into database

Language, abbreviated AQL. AQL is a query language to help users to build patterns that extract structured information from unstructured or semi-structured text. Furthermore, the same component uses Random Forest classifier to rank the relevant documents based on popularity and relevancy scores.

3) Semantic categorizer: 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. [3][4]. 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 catgorization, users can selected health categories of their in-

terest. For example, if someone is interested in prevention related information, then once user clicks on prevention on user interface, only prevention related tweets and news articles will be shown [17] [12].

1.4 Evaluation

Our system extracts relevant information from Twitter for a given user query in near real-time. In our approach, we have addressed three challenges: keyword-based retrieval techniques, search results that may not be real-time, and search results that contain repeated information. To evaluate the system, we took a qualitative approach and compared our system's results set for a user query with the results from Google time-bound search and Twitter search results for the same user query, and conducted blind user study surveys.

In the chapter 2, we will discuss about the prior work on mining information from Twitter, the approaches used to solve the problem, the technologies for finding relevant documents based on a user query, and also discuss a ranking method of the results from Twitter. In chapter 3, we will discuss the process for collecting real-time data, and technologies that are used for collecting the data in real-time. In the following next chapter, we will discuss the implementation of a triple-pattern based mining approach to extract information from tweets and URLs contents. In the final chapter, we will discuss the results and evaluation technique.

Related Work

In our research, some of the key questions we address are: Why do users search health information on Twitter?, Why is Twitter useful for finding health information?, and How to extract health information from tweets?. In this chapter, we will discuss the related work on addressing the above questions. First, we will explain prior work on text mining (information extraction) on Twitter. Second, we will discuss the significance of Twitter in the area of health research and development. Finally, we will discuss the users search behaviour on search engines and social media with respect to seeking health information.

2.1 Mining Twitter information

Traditionally, search engines are a popular platform for finding health information; social media, however, is quickly emerging as the new preferred platform to share and seek this kind of information. The number of tweets shared on Twitter has increased exponentially from past few years. Extracting useful information from Twitter is challenging given its volume, inconsistent writing, and noise. To address Twitter's information overload problem, many researchers worked on various retrieval models such as a user-based tree model, term-based, and pattern-based approaches. Magnani et al. proposed a term-based model for retrieving conversations from microblogs [21]. In this study, authors proposed the concept of conversation retrieval from Twitter which is user-based tree model to retrieve conversations [21]. 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 relevant conversation, the system ranks the relevant conversations based on text relevance, popularity, timeliness, audience, and density features. One limitation of this ranking approach is that the results does not consider the reliability of the information. A Twitter-based social media analytics system, Twitris, uses Spatio-Temporal-Thematic (STT) processing of the Twitter data [27] [18]. However, most researchers favor a term-based extraction model, which is also known as keyword based extraction. The keyword-based model extracts information based on keyword matching of users queries with Twitter messages (tweets). One of the major limitations of this approach is retrieval of the results without considering semantics of the user query.

In the above research, the whole conversation of users are represented as a tree, and their messages and replies are represented as nodes. These conversations are stored in IR engine such as Lucene for indexing the text which can help the system to retrieve the relevant conversation documents based on the query. After finding the relevant conversation, the system ranks the relevant conversations based on text relevance, popularity, timeliness, audience, and density features. The measurement of user popularity is based on the number of followers and the ratio of tweet and replies received. For timeliness, old conversations contain less weight than recent ones. Moreover, the rate of tweets in a conversation shows the level of interest/emotions; a characteristic/trait known as tweet density. In this system, researchers used an information retrieval software package called Lucene [1] to retrieve the relevant conversation, which is again a keyword-based retrieval model. Also, the study has not undertaken the query expansion for extending the results.

Another microblog retrieval framework uses topical features to index documents [19]. The topical features include named entities (person, proper nouns, and events) and phrases. This retrieval framework considers both term-based and pattern-based approaches to retrieve documents. The first step is to extract features and create patterns from microblog

documents and index them into the Lucene. The second step is to extract terms from the query and retrieve the relevant documents based on the search terms. The third step is to expand the query by extracting more relevant terms from the top 10 retrieved documents. The last step is to extract more documents using an expanded query and extract only the top 1,000 relevant documents. This research is also based on a keyword-based retrieval model to extract information. After query expansion the recall is improved. Also, in query expansion only more terms are added to fetch the information, and while the results contain expanded terms, often they are separated very far or there is no relationship between terms. For example, for a user query seeking to know what are the causes of diabetes, after expanding the term diabetes with sugar, the result contains no relationship, e.g., re-educated myself on the evils of sugar and the effect of insulin. Furthermore, this retrieval model has used only microblog messages, and each message is limited in length (tweets are limited to 140 characters). Although limited size messages enable user to read and search information faster, often useful information cannot be written in such a limited length, so people use abbreviations, slang, and other colloquialisms to convey information.

2.2 Usage of Twitter in Healthcare

Many health organisations, the general public, hospitals, and medical professionals are using Twitter to share health information, and to communicate with health consumers. Recently, Twitter has become a venue for the general public, as well as medical professionals for seeking and sharing health information. Patients are using Twitter to find out information on chronic health conditions such as diabetes. According to a USF health survey, people reported that they feel a lot better right after reading Twitter content about diabetes[26].

The World Health Organization used Twitter during the influenza A (H1N1) pandemic for tracking disease and public sentiment. The initial breakout of H1N1 influenza, or swine

flu, was in April 2009. The U.S. Center for Disease Control and Prevention refer to it as ”novel influenza A (H1N1)” or ”2009 H1N1 flu”[9]. In this study, researchers observed that people were not only talking about general information on H1N1 but also sharing countermeasures, consumption-related concerns, treatment-related terms, antiviral medications, etc. Over time the percentage of these influenza-related tweets started to decline rapidly. This research on Twitter shows that whenever the disease occurs, people share a lot of information on Twitter. Tracking the volume of these disease-related tweets can then be useful for retrieving community-level health information.

According to the Spark Report [24], one-third of all hospitals in the US are taking part in social media. Of those hospitals, 64% use Twitter for various purposes such as marketing, patient education, and professional collaboration. The report also indicated that 41% of people said social media would affect their choice of a specific doctor, hospital, or medical facility, and 30% of adults said they were likely to share information about their health and also post about doctors feedback, medical institutes, drugs, and health plans. In this survey on Twitter, 63.2% of the 48 participants reported that they intended to share information about their immediate health status or symptoms and 34.2% wanted to share information or news about a condition. In the USF health survey, 147 people from various age groups and genders participated [26]. Of those surveyed, approximately 25%, 20%, and 5% of people said that they have used social media (Twitter) less than 5 minutes, 10-20 minutes, and more than 2 hours, respectively in order to find out information pertaining to diabetes. The main purpose for using social media was to:

- express their opinions,
- seek information on diabetes,
- have social support from others with similar chronic conditions.

As we have seen so far, researchers and health organizations are focusing on Twitter to analyze data on specific health conditions. At the same time,Twitter has been overloaded

with health information from various sources such as researchers, health organizations, and patients. This has motivated us to investigate Twitter to retrieve useful health information based on a given user query on a health conditions.

2.3 Comparison of Microblog Search and Web Search

Web search engines and microblogging search services are among the most popular tools for seeking and sharing information. The study [29] explored users search behaviour on social media and web search using analysis on large-scale search query logs and supplemental qualitative data. According to the study, people have different intentions for seeking information on Twitter and search engines. While some users search Twitter to find timely and relevant documents (breaking news, real-time content, and popular trends), to read information/posts related to celebrities and influential figures, and to learn about general sentiments and opinions on specific topics, others use search engines to learn about a topic (e.g., facts and navigational information). Many researchers have showed that social media (e.g., Twitter) has become an alternative platform for seeking health information. One of the studies [29] showed that people use social media to search different kind of information as compare to search on Web search engine. In this research, to find preliminary evidence of user intentions with respect to search information on Twitter, the researchers conducted a survey and asked: When you search Twitter, what kind of information are you looking for? A total of 54 Twitter users from Microsoft participated [29]. Participants were allowed to enter freeform text (answer) to the question. Once the responses were collected, researchers labeled each response with multiple categories. In the survey, almost half of participants (49%) reported an interest in searching Twitter for timely information, popular topics, news articles, and events. One-fourth of participants also reported that finding social information was a major intent; over one-third of the participants surveyed used Twitter to find information pertaining to a specific topic. The researchers also observed that Twitter

search results differ from Web search results. On Twitter, results are typically plain text, though they occasionally contain one or more web links. This is in contrast to web searches results, which are algorithmically filtered and presented with links and a short snippet of text. In conclusion, Twitter is useful for finding real-time information which also involves human communication aspects. People include web links in the tweets as reference to the source(s) of the information due to size limitation of the tweets . These web links can be useful for finding relevant information of a users question. In our research we include tweets and web links which are mentioned in a tweet as a data source for finding information.However, extracting useful information from Twitter is difficult due to the information overload.

People use Twitter for seeking and sharing health information on various medical conditions. A study conducted by Choudhury et al., showed that people prefer search engines while seeking information for various sets of medical conditions, and prefer Twitter for sharing information around symptoms of health issues [11]. In the study researchers discovered trends in peoples health activity, the characteristics of that activity, and the information that they seek and share via search engines and Twitter. Researchers have conducted a survey to find out the intent of health searches by people in Web searches and on Twitter. The survey was comprised of 37 questions which were answered by 237 respondents. The results of the survey with respect to user intent and motivation showed that 197 survey participants use search engines, and 40 participants use Twitter to seek health information. The most common motivations for using a search engine were to identify treatment options, to diagnosis a health condition, and to get a general understanding of a health condition or procedure. Whereas on Twitter, people search for information on the most recent events, support or advice on various health conditions, etc. People choose search engines for finding out more information or when they were dissatisfied with medical professionals; some participants also reported that they prefer search engines because of after-hours for doctors. Similarly, people also that mentioned convenience in search on Twitter as a common reason

for seeking health information. They also mentioned that Twitter results consisted of a large variety of relevant information while also allowing for a greater degree of interactivity and personal engagement with the information itself and the information providers. In other words, Twitter enables users to become active participants in the dialogue on health-related information.

In conclusion, microblogging services such as Twitter are used for the seeking as well as sharing of health-oriented information. As we discussed above, when people use Twitter to search for answers to health issues and questions, they are using it primarily for access to timely/time-sensitive information. However, to extract relevant answers is a tedious task, and many times the answers to questions were buried in Twitter because of its real-time nature. Also, the existing Twitter search functionality is restricted to keyword-based retrieval and the search algorithm use only tweets as a data source.

Data Collection and Feature Extraction

In this section, we will discuss the data collection processes, and technologies that are used for collecting the data in real-time. Also, we will discuss the process of feature extraction from the real-time data. For the experiment and evaluation of the proposed research, we have selected diabetes related data as a use-case.

3.1 Data Source

In the ”Related Work” (chapter 2) we discussed the importance of Twitter in healthcare. In this study, we have used tweets (messages shared on Twitter) and URLs content (for URL mentioned in the tweet) as the data sources to extract relevant information for a given user query.In the data collections section, the first we explain, Twitter and its characteristic than will explain how Twitter is selected as a data source. Finally, we will discuss the public API for collecting real-time tweets.

3.1.1 Twitter

Twitter is an online social networking service. It has 500 million users, out of which 284 million are active users [31]. These users are sending and receiving messages are called tweets which fit within the sites 140-character limit [39]. The tweets cover a broad range of topics from political news and product information to healthcare information and are

visibly available to all registered and unregistered users. Due to the size limit of a tweet, it also allows users to share URLs (e.g news articles, breaking news etc.). People post link(s) to provide reference to detailed information. URL can be of any kind, such as an image, article, video, etc. It is also important to know that Twitter uses a short-URL service to make URLs shorter because of the tweet length limit.

Following are the different categories of tweets [28]:

Undirected tweet: A tweet containing no references to others user is called an undirected tweet. The tweet can contain information, status updates, a personal feeling, etc.

Re-tweet: A tweet containing RT in a message is called retweet. If a user is interested in someone's tweet, they can share it with a retweet.

Reply: f users intent is to reply to someones message, he/she can use a reply tweet. If a tweet starts with @username, then it indicates a reply tweet.

Mentions: A tweet just referring to some other username in a tweet by @username but which is not intended to reply is called mention tweet.

On Twitter, to get a users messages, it is required to follow other users. When users choose to follow other users, the subscribers are known as followers. Whenever a user posts a new tweet on their profile, it immediately appears to all the followers. All the tweets appear in reverse chronological order on a Twitter page, users can be updated with information in real-time. In addition to that, users can also select a location to be assigned to the tweet[28].

Although Twitter does not provide a way to send group messages, the users can still send messages their own way using hashtags (a hashtag is a word or phrase prefixed with a # sign and lacking spacing between individual words). Hashtags have become very popular for retrieving trending topics. A trending topic is a word, phrase, or topic that is discussed more frequently than others. Also, the top 10 trending topics are shared on Twitter's home-pageconsidered the full set of tweets. A Twitter user can post a message on twitter with the #, which helps to make that topic popular [28]. The tweets contain huge amount of information.The information is not only in a tweet but also present in shared URLs which

may appear in a tweet.

3.1.2 Twitter Streaming API

Twitter offers a set of streaming APIs: public streaming, user streaming, and site streaming [32]. We have used the public Twitter streaming API to collect health-related tweets. This public streaming API establishes a persistent HTTP connection to the Twitter service [32]. Before establishing a connection, Twitters servers requires authentication. Once the connection is established, the application starts getting a feed of tweets. The server sends its response in the JSON format. JSON is stands for JavaScript Object Notation; it is a simple and easy to parse format for describing structured data. However, Twitters server has limitations. It sends the response in a block, and its size is allowed to be 1,500 bytes. When the connection is idle and there is no other data to send, the streaming API sends an empty signal every 30 seconds to keep the connection alive [32].

To get filtered tweets, the streaming API requires keywords to track public tweets. These keywords could be a word or phrases and should be separated by a comma. Twitters server matches the keywords with the tweet. If matches are found, then it pushes the tweets to the client. In addition, Twitters server returns metadata associated with the tweets such as display url, hashtags, temporal, and location information.

Many languages have implemented the Twitter Streaming API, such as Java. In our system we have used the Java-based Twitter streaming API library. The library has a Configuration class which provides developers with the ability to pass authentication details and keywords. Next, this configuration object is passed to Twitters server to verify the user. The keywords help the Twitter server to filter out the tweets. With the help of the TwitterStreamFactory class, the client can make a connection and pass the configuration to the server. In the next section we will discuss the method of collecting the keywords.

```
1 Configuration twitterConfig = new ConfigurationBuilder();
2 TwitterStreamFactory fact = new TwitterStreamFactory(twitterConfig);
3 FilterQuery filterQuery = new FilterQuery();
4 filterQuery.track(this.keywords);
5 twitterStream.filter(filterQuery);
```

3.1.3 Unified Medical Language System (UMLS)

Twitters streaming API requires keywords to filter out health-related tweets. We have used UMLS-Metathesaurus to collect authentic and reliable keywords. UMLS is a system that brings together many health and biomedical vocabularies [23]. These vocabularies can be useful to developers for creating applications related to classification tools of various medical records, creating dictionaries, etc. [23]. Developers also use vocabularies in data mining for health-related data. It can also be useful to make a knowledge base for various computer science applications using medical terms, drugs names, etc. [23].

3.2 Tweet Retrieval and Feature Extraction

Here, we will discuss the process of collecting the tweets and how the feature extraction process works in real-time. We have used Apache Storm to collect real-time tweets and to perform real-time computations. Apache storm uses the public streaming API to collect tweets. In this section, we will also discuss the architecture.

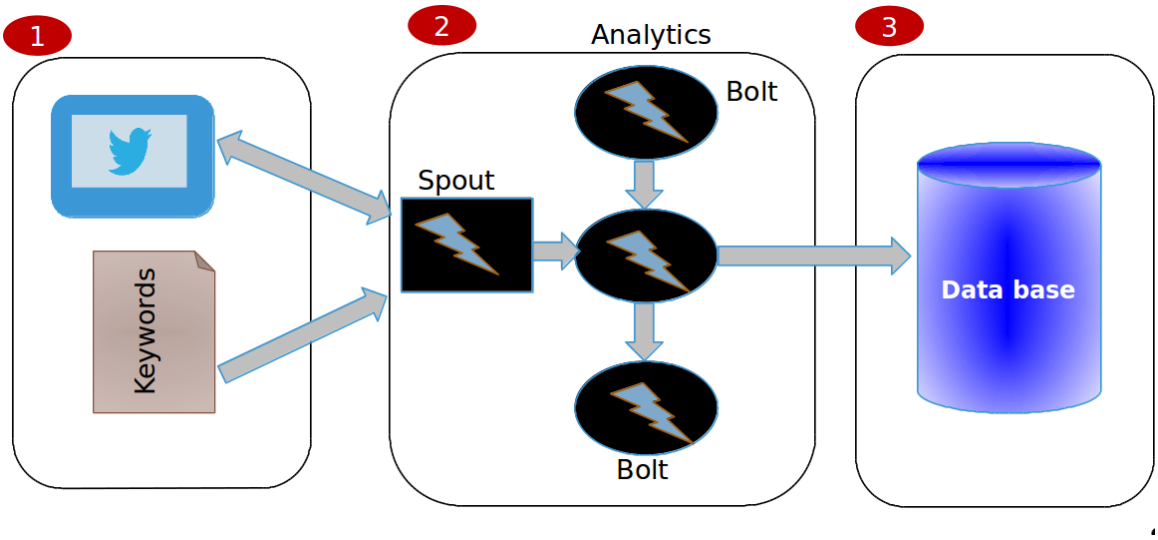


Figure 3.1: Tweet retrieval and feature extraction architecture

3.2.1 Architecture

The Figure 3.1 is an architecture diagram of tweet retrieval and feature extraction. Apache Storm is the main component which extract tweets using the public Twitter streaming API while also performing computations. In this diagram, the first part is Twitter as data source, the second part is an analytic component (spout and bolt), and the final part is a database. We will explain Apache storm and its components in the following subsections.

3.2.2 Apache Storm

Apache storm is free, open source software used for real-time, distributed computing [4]. It is similar to Hadoop for processing batch process data. It is a reliable process for computing streaming of data. To do real-time computation in Storm, we have to create a ”topology.” A topology is a network of computation. Each node in a network contains logic for real-time computation. There are two types of components in a topology: bolt and spout. A spout is a source of stream data. It reads the data from any source (e.g Twitter, Kafka, etc.), and converts the data into tuples. These tuples pass to bolts one-by-one according to the topology structure. A bolt consumes the tuple and performs computation logic. Once the

bolt is finished, it sends an acknowledgment to the parent bolt and passes the tuple object to the next bolt. There are many operations performed on tweets such as filtering tweets, streaming aggregations, streaming joins, talk to databases, and more [5].

Once the network of spouts and bolts are packaged into a ”topology”, then programmer submits it to storm clusters for execution [5]. Storm clusters are superficially similar to Hadoop clusters. There are two kinds of nodes in storm clusters: the master node and worker nodes. The master node is called ”Nimbus”, which is similar to Hadoop's Jobtracker node [5]. It is responsible for distributing code and assigning the tasks to machines. It is also responsible for monitoring for failure. Similarly, worker nodes are called ”Supervisors”. They are similar to Hadoop's Tasktracker nodes. The actual task is to perform the execution of topology [3]. A zookeeper cluster coordinates all process between the Nimbus and Supervisor nodes [5]. Figure 3.2 below shows the topology network.

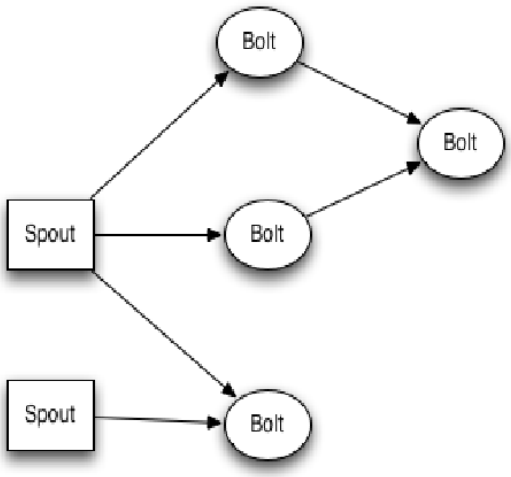


Figure 3.2: Storm's Bolt and Spot topology

3.2.3 Feature Extraction

A tweet has many features, such as text, short url, latitude and longitude, retweet count, etc. All these features can be helpful for discovering useful information. To extract all

theses features from tweets in real-time, we have used Apache Storm components (spout and bolt). This process is also known as a pre-processing analytic pipeline (Figure 3.3), because the extracted features and data help to pattern the extraction module. This dataset will be required for extracting useful information based on a user query. A spout use the Streaming API to crawl real-time tweets. The bolts contain computation logic to perform feature extraction logic in real-time. Once all computing bolts are finished, the final bolt will save the data into the database. The first bolt is a filter bolt to identify the language of a tweet and allow only English tweets.

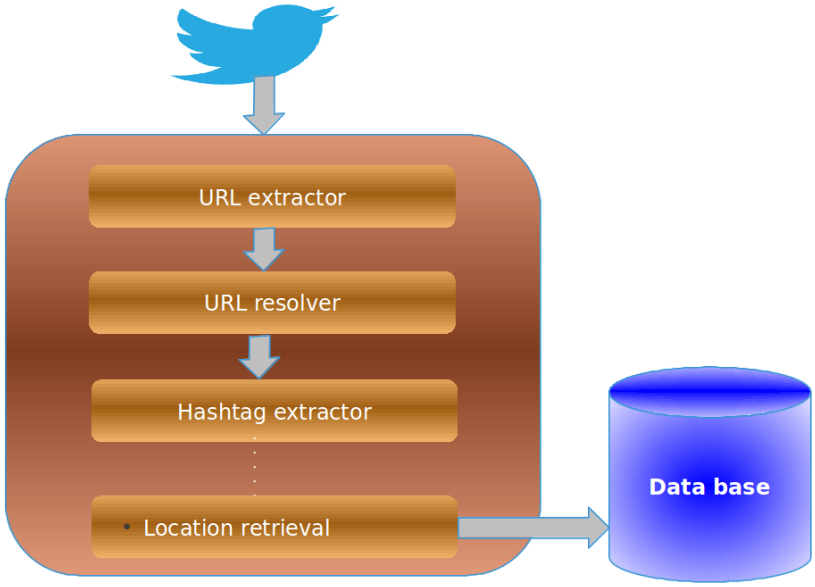


Figure 3.3: Feature extraction pipeline

The Table 1.1 shows a brief description about each bolts computation logic. Later in this section we will discuss this in more detail.

Twitter Spout: We have only one spout in an analytic component, called the Twitter spout, which reads tuples from an external source (e.g, Twitter) and emits them to the topology. The Twitter spout uses the Twitter streaming API to collect a stream of tweets, and then it provides the status. A status object contains all the information such as text, URLs, screen name, user details, media objects, metadata information, etc. Once the Twitter server

Table 3.1: Apache storm pipeline- Feature extraction components

Component name	Description
Language identification	For filtering out non-English tweets
Crawler	It crawls the real time tweets from Twitter based on the keywords
Hashtag retrieval	It is used for retrieving hashtags from the tweets
Informative analysis	Analyzing how informative is a tweet
URLs resolver	Expanding the URLs(weblinks in the tweets) from short to its original form
URLs extractor	Extract URLs(weblinks in the tweets) from tweets
Location retrieval	To retrieve geo-coordinates of the tweets
DBHandler	To save all the extracted features and tweet object into database

sends a status object, it is immediately queued into the LinkedBlockingQueue (Java class) and deques one-by-one to the bolt. To implement the spout, a user-defined class should implement the IRichSpout interface.

Hashtag Retrieval: This is a bolt class which implements the IRichBolt interface. The main function of this class is to read the status object from the tuple and extract hashtags from text. We have used regular expression to extract hashtags from text.

Language Identification: This is also a bolt class, the main function of this is to identify the language of tweet. It passes the tweet to the remaining bolts for computation only if a tweet is in English. We have used one of the lang attribute of status object (Twitters Streaming API object) to identify the language. Sometimes the langs value is empty, so to determine the language we use the Apache library from the given text.

Location Retrieval: This is used for finding out the location of tweet. The first way is to extract geolocation coordinates from tweets if it is available. Alternate approach is, to take the location name from geo-coordinates through OpenStreetLocator Java API. The final way is to get the location from users profile locationonly if geo-coordinates are not tagged in tweet.

Informative Analysis: It is used to identify the informativeness of tweets. We are using tweets features, such as its length, URL, reply count, retweet count, and hashtags, to calculate the informativeness. Each feature is assigned some weight, and the sum all the features weights decides the informativeness score.

URL Extractor: This bolt is to extract URLs from the tweets. We have used the Java regular expression for extracting URLs.

URL Resolver: The aim of this bolt is to expand the short URLs. All the extracted URLs the URLs extracted were previously short form.

DBHandler: Once all features are extracted, they are stored in the database. We are using a MySQL database.

3.2.4 Dataset

Once the analytic pipeline is finished, the final bolt, DBHandler, is called and it stores all the extracted features. We have categorised the features into two parts: Tweet metadata and URL metadata. The first is tweet metadata in which we store tweet-related metadata, such as text, retweet count, retweet, temporal and location information, etc. The final category is URL metadata, in which we store the short URL, the expanded URL, URL content, etc. Later, a pattern extractor uses this metadata to extract a social media ranking (Twitter share count, Facebook likes and comments count) to rank URLs. We will discuss this in the next chapter.

Following tables (3.2, 3.3) are a description of categories and their metadata:

Table 3.2: Tweet metadata

Meta-data	Description
tweet_text	This field contains a tweet of any user
retweet_count	It contains re-tweet count of a tweet
is_retweet	It contains true if retweeted, otherwise false
total_informative_score	It contains the value of analysis of tweets
loc_latitude	A geo-coordinate of a user or tweet (only latitude)
loc_longitude	A geo-coordinate of a user or tweet (only longititude)
country	A country name of a user
state	A state name of a user's country
hash_tag	All hashtag which is presents in a tweet
disease_category	It contains the value of tweet disease category

Table 3.3: URL's metadata (present in a tweet)

Meta-data	Description
raw_url	In this field, we store the a URL present in a tweet
long_url	This field contains a URL in the original form
url_content	This field contains the content of a URL
facebook_total_count	It contains the facebook's URL total count (share+like count)
facebook_like_count	It contains the facebook's URL like counts
facebook_share_count	It contains the facebook's URL share counts
twitter_count	It contains the twitter's URL share counts
google_domain_count	This field contains the URL's Google domain page rank

Extraction of Relevant Documents

The aim of this thesis is to extract relevant documents based on a given user query. Here, a key challenge is extracting the information from unstructured data in near real-time. In this section, we will discuss the implementation of a triple-pattern based mining approach from tweets and URLs contents.

4.1 Introduction

In the previous chapter, we discussed the process of collecting real-time tweets and extracting features. Once the features are extracted, the information extractor module collects all the stored tweets and their features for extracting relevant information using a triple-pattern based technique. We extract information at an interval of every six hours. To extract relevant information and/or documents, we have used the IBM text analytic Annotated Query Language, also known as AQL. AQL is a query language to help developers to build queries that extract structured information from unstructured or semi-structured text [16]. We have used an AQL to construct a triple-pattern. The triple-pattern (subject, predicate, and object) is defined in the initial question. We have divided users questions into two categories: static and dynamic. Static questions are the most frequently asked questions collected from different sources. The dynamic questions are typed by the user on the fly, which is not the case with static queries. Once the results are extracted, we use a Naive Bayes classifier to rank the results based on the popularity and relevancy score of the URLs. This module is

implemented in Apache Hadoop to handle a large dataset.

The information extractor has three modules: a URL extractor, a social media rank extractor, and a triple-pattern extractor. These modules are invoked at an interval of six hours. The first module is a URL extractor which collects URLs from the database and extract the URLs content. We used the Jsoup Java library for parsing the HTML document to extract content. The second module is a social media extractor in which we are drawing the social media (Facebook, LinkedIn, Google domain rank) ranking of URLs, which helps to rank the final results. The final module is a triple-pattern extractor which extracts patterns from the tweets and URLs content using the AQL queries. All these modules are implemented inside Apache Hadoop, we have used a JobController (Hadoop Java class) to execute all the modules one-by-one. In the next sections, we will introduce the details of all modules and the system architecture.

Architecture Diagram

Following is the Figure 4.1 of an architecture diagram our system:

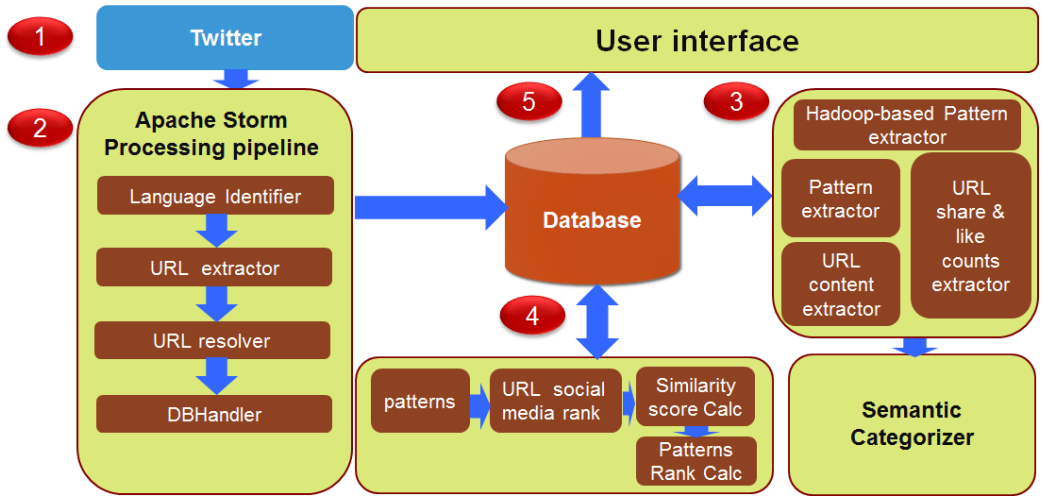


Figure 4.1: Architecture diagram of Social Health Signals platform: It has three major components- an Apache storm pipeline, a Hadoop based pattern extractor, and Semantic categorizer.

We have developed a system to enable a user to either to dynamically ask questions or select frequently asked questions. The first component is a real-time analytic component which we discussed in the third chapter. The second and third components will be discussed later in this chapter.

Question Categories

We have categorised the users questions into two parts: static and dynamic. The static questions are most frequently asked questions; they are collected from various websites such as WebMD, Mayo Clinic, etc. These static questions are related to diabetes. A user can also ask questions dynamically. To extract relevant documents based on the questions, we used the IBM text analytic tool is AQL (Annotation Query Language). To build an AQL query, we have to extract subjects, objects, and predicates from the questions. However, the processes of extracting relevant documents based on static and dynamic questions is different. In the case of static questions, AQL queries are inbuilt queries (developed by the developer) and extract documents every six hours, while in the case of dynamic questions, AQL queries are built on the fly and extract documents at that moment.

The first category is static questions in which the queries are executed inside the Apache Hadoop environment. We extract the synonyms of each token (if the token is not a medical term) in the query using WordNet [40]. Also, we used UMLS[23] to get related keywords for the tokens only if the tokens are medical terms, such as diabetes, etc. AQL provides the feature to build dictionaries of each token, which helps to expand the query and extends the results.

The second category is dynamic queries in which question are typed by the user on the fly; these queries are not available in the database. To extract the relevant documents from dynamic queries requires the same approach (static question), but the process for extracting and building an AQL query happen dynamically. The first step is to extract triples using the Stanford Parser. We used WordNet as a thesaurus to help users find synonyms.

4.2 Apache Hadoop

We have used Apache Hadoop to build the information extractor. Apache Hadoop is an open-source software framework written in Java. It is a set of algorithms for distributing computing and the storage of very large data. There are two core parts of Apache Hadoop:Hadoop Distributed File System (HDFS) and MapReduce. HDFS is used for storing data in distributed manner and MapReduce is used for processing [33].

The HDFS is a file system for distributing data on clusters. It is responsible for storing, deleting, and governing the availability of data, just as in other file systems. It splits data into large blocks (default 64MB or 128MB) and distributes the blocks amongst the nodes into clusters [33]. An HDFS file system has a master/slave architecture. A master node is a NameNode, and a slave node is a DataNode. A master node is responsible for regulating file access by clients and manages the file system namespace [2]. In addition to being responsible for reading and writing requests from the file system, the DataNodes also perform block creation, deletion, and replication upon instruction from the NameNode [2].

A TaskTracker accepts tasksMap, Reduce and Shuffle operationsand then performs computation on data. The JobTracker talks to the NameNode to determine the location of the data and TaskTracker to submit the task. To process the data, Hadoop Map/Reduce fetches data from the file system and performs the computation. Also, Hadoop has many wrapper libraries to get data from different sources such as a MySQL database, NoSQL, etc. In our case, we connect to a MySQL database to get all the data for computation.

4.3 Information Extraction

Information extraction is a task to extract structured information from unstructured and/or semi-structured sources automatically. We have implemented an information extraction module inside the Apache Hadoop realm; it consists three sub-modules: URLs content

extractor, social media extractor, and a pattern extractor. These modules are compromised in one job. This job will be executed at an interval of every six hours.

The algorithms are implemented inside the Map-Reduce framework. This Hadoop application connects to the MySQL database and passes the data to the mappers. A mapper/reducer performs computation on the data and either inserts or updates the results. We have created a chain of Mappers/Reducer to perform a series of operations (three modules). The dependencies of these jobs rely on their configuration, so the jobs can be run in parallel or as a series. However, in our case we invoke them one-by-one (a series of operations).

1) Create two JobConf objects and set all the parameters

2) Then create two Job objects with jobconfs, Job job2=new Job(jobconf2);

3) Using the jobControl object, you specify the job dependencies, and then run the jobs: JobControl jbcntrl=new JobControl(jbcntrl);

In this section we will discuss implementation of modules and their algorithms and also discuss the technologies which are used to implement them.

4.3.1 Extraction of URL Content

This is the first module to execute and it aims to extract content from the URLs. To extract content from the URLs, we have used an HTML parser library called Jsoup (Java library). It provides APIs for extracting and manipulating data using DOM traversal or CSS selectors. It has implemented the WHATWG HTML5 specification for parsing HTML. It also helps to manipulate the HTML elements, attributes, and text. In this instance, however, we use HTML parser to parse the URLs content. The last part of this step is to store extracted content from the URLs back in the database.

Following below are the steps to retrieve content:

1) Get the tweet object from the database

2) Retrieve the long URL, which is the resolved URL, from a tweet object

3) Convert the URL to a Java URL object

4) Pass the URL object to the Jsoup parser

5) The Jsoup parser retrieves the content from the URL and parses the HTML

4.3.2 Social Medias URL Shares and Like Counts

People share URLs on social media for detailed information. People also click on like buttons to show positive feelings towards and approval of shared links. These shares and like counts show the popularity of URLs on social media. In our case, we utilized URLs shares and likes counts for ranking the results. To extract the shares and likes counts (including Facebook shares, Facebook likes count, Twitter shares count, and Google domain pagerank), we used public social media APIs; it should be noted that the response of the public APIs are in JSON format. The Mapper/Reducer job stores the result back in the database. Later, these counts will be used to rank the final results.

Following are the steps involved in this module:

1) Get the tweet object from database

2) Retrieve long URL, which is the resolved URL, from tweet object

3) Pass the URL to Facebook, Twitter, and Google APIs to get shares and like counts

4) These APIs return a JSON object

5) Extract shares and like counts from JSON object

4.4 Extraction of Triple-Pattern

This module is a very important module for extracting relevant documents based on an AQL query. The relevant documents are triple-patterns found in the URLs content and tweets. This is the last module to be executed. In this section we explain the triple, AQL query, and the process of extracting documents from URL content and tweets.

4.4.1 Triple

A triple-pattern consists of three parts: subject, predicate, object. The subject and object are a noun or noun phrase; similarly, a predicate is a verb, verb phrase, noun or noun phrase. Triple-patterns are written in the form of a subjectpredicateobject expression or a whitespace-separated list. RDF (Resource Description Framework) is part of the family of the World Wide Web Consortium (W3C) standards and uses the triple-patterns format to store results. We have used this triple-pattern format to get the answers (relevant documents) to a users question.

4.4.2 Annotation Query Language (AQL)

Annotation Query Language (AQL) is a language used for building queries that pulls structured information from unstructured or semi-structured text [16]. It is one of the components in InfoSphere BigInsights Text Analytics [16]. To execute the AQL queries, first, we need to compile the input top-level AQL file or string to the compiled operator graph (AOG). SystemT translates AQL statements into an algebraic expression called an AOG. SystemT is a high-performance run-time and uses optimized execution plans. The execution plans are extractors with rule semantics that obtain structured information from unstructured documents. IBM provides Java APIs to perform all these operations.

Also, an AQL query provides a faculty/tool called a dictionary, which we use to define a set of terms that will identify matching words or phrases in the input text. There are two types of dictionaries: internal and external. Internal dictionaries contain a synonym that is specified in the AQL, whereas external dictionaries contain a synonym that is not specified in the AQL. Instead, it is supplied when the compiled extractor is run. We have used this feature to expand the query.

1) Get the tweet object from database

2) Retrieve URL content and tweet from tweet object

3) Pass an AOG file and content to SystemT

4) System returns a set of structured information

5) Iterate the results set (Java class TupleList) and save into the database

Ranking of Results

As discussed in the previous chapter, the results are triple-patterns which are extracted from tweets and URL contents. In this chapter, we will explain the ranking of results. To rank the results, we have evaluated various ranking algorithms. We will show the evaluation of each algorithm and feature extraction in following sections of this chapter. Over the past decade, user content has increased exponentially since the emergence of social media. Unfortunately, not all content is of good quality and the amount of poor quality content is quite high. The presence of both kinds of content (good and bad quality content) on social media has led to users engaging with search engines to retrieve useful information for queries. In addition to simply receiving answers, users want the results to be high-quality and well-ordered. Popular Web search engines are focused on ranking algorithms to order the results. The majority of algorithms in place are machine learning algorithms which rank the results based on popularity, relevancy, etc.

Machine learning algorithms create models from input data and use those models to make predictions [36]. We have tested many machine learning algorithms for this work and evaluated the results. Based on an evaluation matrix, we have chosen the Random Forest algorithm.

Random forests are an ensemble classifier that operate by constructing a multitude of decision trees. Decision trees that are grown very deep tend to overfit (i.e a model starts to ”memorize” training data instead of ”learning” from the training data) their training sets because they have low bias but very high variance. However, the Random Forests algorithm

has a way to average down multiple decision trees (i.e., train on different parts of the same training set) with the goal of reducing the variance [37]. We have selected 100 trees for classifying and for ranking the results. Additionally, we have selected the default values for all other parameters.

To rank the results, the algorithm requires the building of a model from input training data. Our model is based on social media features such as the likes, share counts, and a similarity score. Each training datum is a vector of features. These social media features measure the popularity of a web link from which the pattern extracted, and the similarity score measures the relevancy of the extracted pattern to the user question. To train this algorithm, we have labeled the data based on similarity, popularity, and relevancy of web links.

An initial step in our application is to create a new set of features to facilitate learning. In our application, there are two sets of features: popularity and relevancy. The popularity set has the share and like counts of web URLs on various online social media platforms. Similarly, to know how the extracted patterns are relevant to the users question, we have used a string similarity algorithm.

5.1 Features Selection

A feature is an individually measurable property for a machine learning algorithm. The selection of the features is a crucial step. Therefore, in this section we will discuss the process of feature extraction for ranking.

5.1.1 Popularity Features

Social media is a platform that allows people to create, share, like or exchange information. So this share and like information is used to find the popularity of the source of results.

As there is a limitation to the length of posts, when people wish to share more detailed information they often insert a link to share with their friends, readers, followers, etc. These shares and like counts are the measures of the popularity of web link. In our research, we used this as one of the features in our algorithm. We have used Facebook shares, like count, Twitter share counts, and Googles URL domain rank as the features. We have used the respective social media APIs to extract the shares, likes, and URL domain ranks.

5.1.2 Relevancy Features

In information retrieval, similarity algorithms are used to match the similarity of extracted information. This feature is used for measuring the relevancy of documents based on a user query. There are many approaches to solving this problem. We have used two algorithms: (1) a vector space model and (2) a bag-of-words model. In the first model we have used the TF-IDF (term frequency-inverse document frequency) algorithm. In the second model we have used a Jaccard coefficient algorithm.

Jaccard coefficient

It is a commonly used measure of the overlap of two sets. The set is a collection of unordered set of words extracted from documents. However, this model does not consider the frequency and rare occurrence of words [35].

$$JC = \frac{(AnB)}{(AuB)} \tag{5.1}$$

TF-IDF (Term Frequency-Inverse Document Frequency)

It is a numerical statistic often used as a weighting factor in information retrieval and text mining. It has two parts, term frequency and inverse document frequency [38] [30].
TF (tTerm frequency) is used to determine which document is most relevant to the query.

The importance of a document refers to how often a term occurs in a document, that is, the total number of word occurrences in the documents. As the documents are different in length, it is possible that a term appear more times in longer documents than shorter ones [38] [30].

$$TF(t) = \frac{Numberoftimestermtappearsinadocument}{Totalnumberoftermsinthedocument} \tag{5.2}$$

Inverse document frequency (IDF): There is a term in the document with an extremely high frequency, such as "the," which does not give a significant enough weight to the meaning of the document to warrant being counted as heavily as more rare terms. These high frequency but low meaning terms are decreased in weight down while the less frequent terms are scaled up. The words which end up getting lower weights include terms such as is, are, the, etc [30].

$$DF(t) = log_e(\frac{Totalnumberofdocuments}{Numberofdocumentsmatchingterm}) \tag{5.3}$$

We collect the extracted patterns from the database at an interval of six hours and then calculate the TF-IDF score based on the user query. Finally, the TF-IDF score is stored back in the database.

Evaluation of ranking algorithm

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, relevance, 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 page rank (filtration criteria is page rank 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 Random Forest algorithm based on an evaluation matrix (Normalized discounted cumulative gain). Please see the Figure 5.1 for comparison of ranking algorithms based precision and recall.

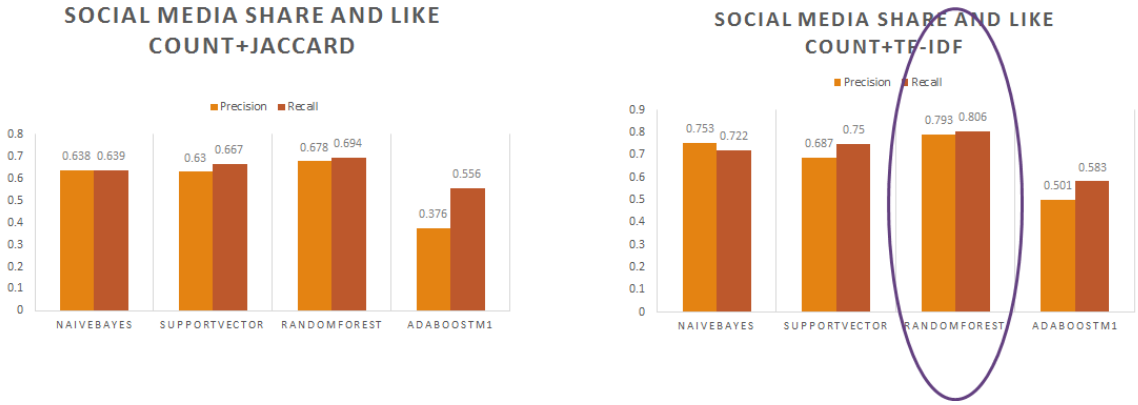


Figure 5.1: Comparison of ranking algorithms based precision and recall

Result and Evaluation

In this chapter we will discuss the results of our research. As our research is focused on extracting near real-time health information based on users search queries, we have made the decision to compare our systems results with those of existing real-time search engines as well as search engines that allow for a search within a specified/custom date range. We have chosen a Twitter search and a Google time-bound search as the points of comparison. Twitter is a very popular platform for seeking latest health information. Similar to Twitter, Googles search is also popular for finding out latest health information, such as news articles, etc., but to find the latest information through Google requires the use of a custom date filter option. In our research on real-time health information, we conducted a survey which takes into account three questions dealing with the chronic disease diabetes.

1) How to control diabetes?

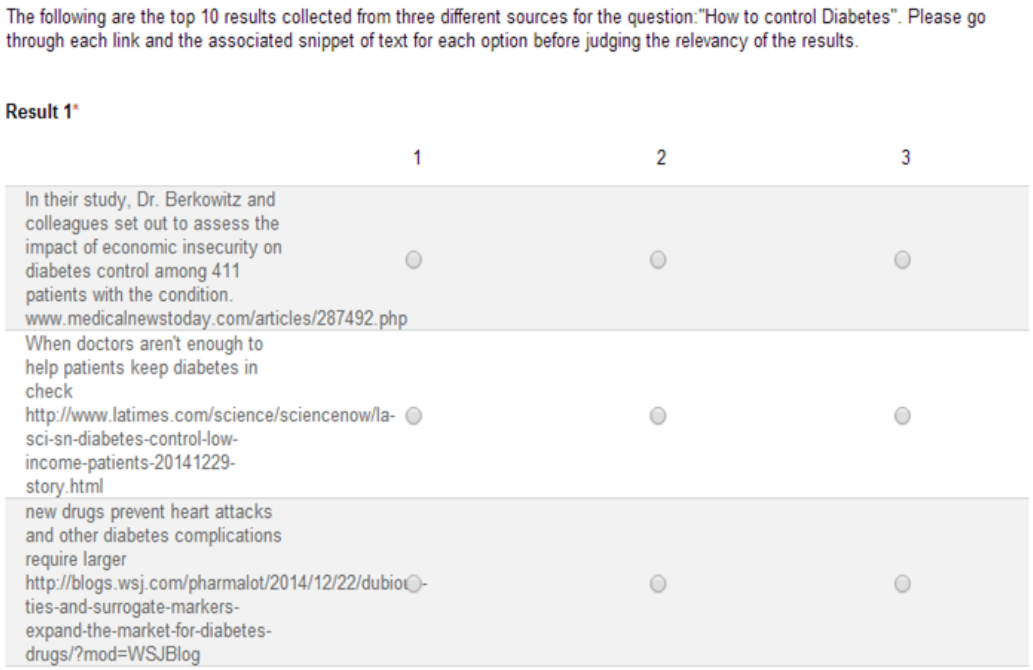
2) What are the causes of diabetes?

3) What are the symptoms of diabetes?

6.1 Survey

While people perceive search engines as providers of higher quality health information relative to other internet sources, their desire for real-time information and for engagement with content providers often leads them to seek out other sources of information. Given

that there is room for improvement in the existing technology, our researchs primary intent is to show that seeking near real-time health information without using a keyword based retrieval approach. We have adopted a survey method which asks participants to judge the results to questions asked on multiple platforms. In our research, we conduct three surveys for each question. Each survey consists of the top 10 results of a Google time-bound search, a Twitter search, and the results of Social Health Signal (our system) presented together in sets. There are ten sets in each survey and a set consists of three results, one result from each of the three sources. For example, the first set consists of the top ranked result of each source. Similarly, the second set consists of the second most highly ranked result from each source. Users judge each document in a set on a scale from 1 to 3 (1-not good, 2-good, and 3-very good). Please see Figure 6.1.


The following are the top 10 results collected from three different sources for the question:"How to control Diabetes". Please go through each link and the associated snippet of text for each option before judging the relevancy of the results.

Result 1*

	1	2	3
In their study, Dr. Berkowitz and colleagues set out to assess the impact of economic insecurity on diabetes control among 411 patients with the condition. www.medicalnewstoday.com/articles/287492.php			
When doctors aren't enough to help patients keep diabetes in check http://www.latimes.com/science/sciencenow/la-sci-sn-diabetes-control-low-income-patients-20141229-story.html			
new drugs prevent heart attacks and other diabetes complications require larger http://blogs.wsj.com/pharmalot/2014/12/22/dubio-ties-and-surrogate-markers-expand-the-market-for-diabetes-drugs/?mod=WSJBlog			

Figure 6.1: A survey example for “How to control diabetes? ”

Since a Google time-bound search provides many custom dates (anytime within the past month, week, 24 hours, hour, etc.), we have chosen the past 24 hours as the custom time frame of all the queries to mimic the abilities of a real-time search engine. Similarly,

we have considered top 10 most popular results of 24 hours of tweetssometimes the search results may not be latest (Figure 6.2) . In our case, we have observed the collected data hours and found that 6 hours worth of data or a 6-hour time frame provides a sufficient quantity and quality of results for finding useful information [Table 6.1]. In our surveys, a total 15 participants judged the results. Upon completion, we found that participants ranked our systems result higher (very good) than those all of the other sources studied. We will discuss our observations and results in following sections of the chapter.

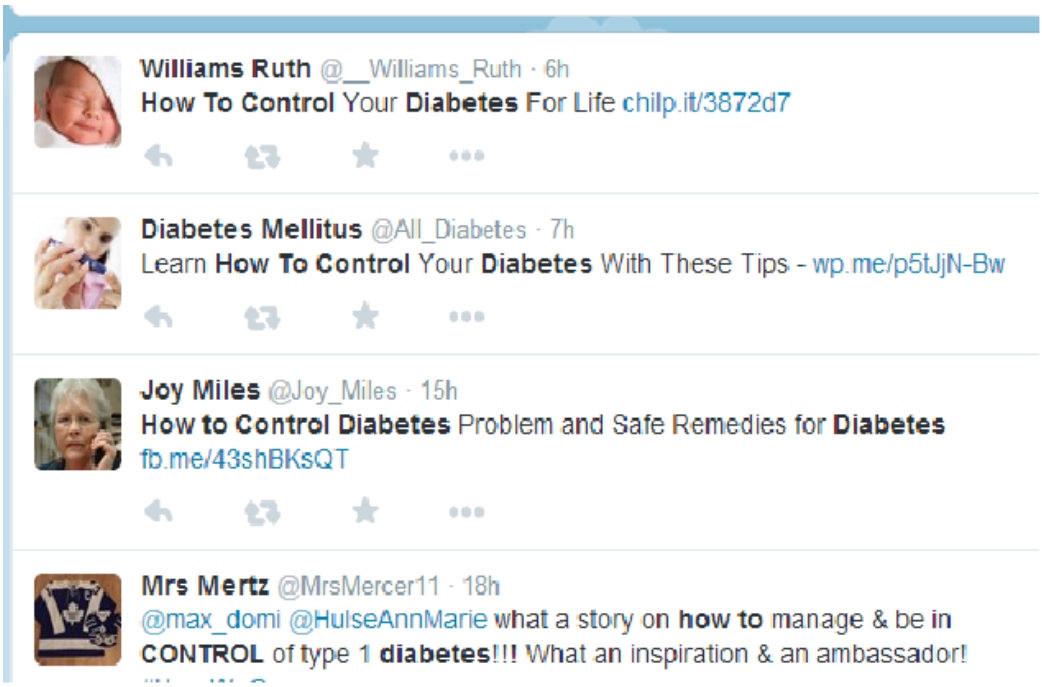


Figure 6.2: Twitter filters search results based on keywords, and the results are not latest (e.g., 6h, 7h, 15h)

6.2 Results

To evaluate the system, we conducted blind surveys in which participants did not know which sources corresponded with which results. We also observed, apart from the quality of the information, that people ranked as bad quality those results which did not contain

any web link, and people ranked as bad duplicated results (different web links).

Query1 (Survey 1):

In the first survey, we ask participants ”How can diabetes be controlled?” Participants were required to rank the answers by selecting ”bad,” ”good,” and ”very-good” (1, 2, and 3). The answers to this question are searched from Social Health Signals, Twitter, and Google. Upon completion of the first survey, 50% of users ranked the quality of our results as ”very good,” whereas the results for a Google time-bound and Twitter search were 10% and 40% respectively. We have observed that a low percentage of participants ranked the Google time-bound results as ”very good” due, presumably, to the fact that results were dominated by breaking news. Similarly, there is a low percentage of results classified as very good for Twitters search due to repetitiveness. Furthermore, 40% of users classified our systems results as ”good,” while the percentages were 40% and 50% for a Google time-bound and a Twitter respectively. In the case the label good, all the percentages of all the sources are similar. Also, the percentages of participants who ranked the quality of as ”bad” are: 50% for a Google time-bound search, 10% for a Twitter search, and 10% for our results.

Query2 (Survey 2):

In the second survey we ask ”What are the causes of diabetes?” Similar to the first survey, participants respond on a three-point scale from 1 to 3 (1-”bad,” 2-”good,” and 3-”very-good”). Respondents agree highly (60%) that health information available via our system is of a high quality (”very good”) as compared to a Google time-bound and a Twitter search, which are 50% and 10% respectively. The increased percentage of results designated very good is due to the fact that, here, a Google time-bound searchs results do not have repetitive information. Similarly, the percentages of participants who classify results as ”good” are 10% for a Google time-bound search, 60% for a Twitter search, and 30% for our system. This time, the highest percentage of people ranked Twitters search results as ”good”. However, in our case, the most frequent classification/rankings are very good and good. Furthermore, the Google time-bound search results were ranked ”bad” by

40% of the participants, which is similar to survey 1. However, in the case of our system, 10% of participants ranked the results ”bad” compared to a Google time-bound search (40%) and a Twitter search (30%).

Query3 (Survey 3):

In the final survey, we ask the users to judge the top 10 results for the question, What are the symptoms of diabetes? The respondents rankings for ”very good” are 10% for a Google time-bound search, 40% for a Twitter search, and 50% for our results. Again, our results rank very good at a higher percentage than that of any other of the other sources. With respect to the good label/ranking, 70% for a Google time-bound search, 30% for a Twitter search, and 30% for our system. For the bad label, the results break down to 30% for Twitter search results and 20% for both a Google time-bound search and our systems results.

The following Tables (6.1, 6.2, and 6.3) show the percentages of all the surveys results:

Table 6.1: Survey results of a “high quality ”only

Query	**Google Time-bound**	**Twitter**	**SHS(Social Health Signals)**
Query 1	10%	40%	50%
Query 2	50%	10%	60%
Query 3	10%	40%	50%

Table 6.2: Survey results of a “good quality ”only

Query	**Google Time-bound**	**Twitter**	**SHS(Social Health Signal)**
Query 1	40%	50%	40%
Query 2	10%	60%	30%
Query 3	70%	30%	30%

Table 6.3: Survey results of a “bad quality ”only

Query	Google Time-bound	Twitter	SHS(Social Health Signals)
Query 1	50%	10%	10%
Query 2	40%	30%	10%
Query 3	20%	30%	20%

6.3 Evaluation Matrix

MAP@k vs nDCG@k:

The two types of evaluation metrics for ranking are binary relevance and multi-level relevancy. We measure the objective performance of our system using nDCG@K. The ranking algorithms are often evaluated using information retrieval measures such as Normalized Discounted Cumulative Gain (nDCG) and Mean Average Precision (MAP). Mean Average Precision for a set of queries is the mean of the average precision scores for each query. Precision is the amount of retrieved documents that are relevant to the user's query.

Formula for MAP [34]:

$$MAP = \frac{\sum_{q=1}^{Q}(AvgP(q))}{Q} \tag{6.1}$$

NDCG (Normal Discounted Cumulative Gain):

TThe nDCG is a ranking metric. It predicts a list of sorted documents, and then compares it with a list of relevant documents. Its values vary from 0.0 to 1.0, and 1.0 represents the ideal ranking. Also, the nDCG metric is commonly used to measure the performance of search engines. In nDCG, the documents, which are highly relevant, are more valuable when they appear on top in a search result list [34] .

CG (Cumulative Gain): Cumulative Gain (CG) is the prior version of DCG and does not consider the position of a result set [34].

$$CG_p = \sum_{i=1}^{P} rel_i \tag{6.2}$$

DCG: In DCG, the highly relevant documents appearing lower in a search result list should be penalized. It reduces the graded relevance value by a logarithmic factor to the position of the result [34].

$$DCG_p = rel_1 + \sum_{i=2}^{P} \frac{rel_i}{log_2(i))} \tag{6.3}$$

nDCG: Search result lists vary on different search engines for the same query. To compare the performance of different search engines, with consideration for different sets of documents search results for the same query, this cannot be achieved using DCG alone. Therefore, the cumulative gain should be normalized across queries [34].

$$nDCG_p = \frac{DCG_p}{IDCG_p} \tag{6.4}$$

6.4 Evaluation

An MAP ranking evaluation is based on a binary relevance. This means a document can be either relevant or irrelevant. However, in our case, we have used NDCG@k because we have considered multiple relevancy to judge the results, and NDCG@k considers multiple relevancy scores. Upon the completion of all the surveys, we calculated the average ranking for all the participants in every survey for an NDCG evaluation matrix.

Average score calculation formula [22]:

$$Rank = \frac{(TotalCountof3s * 3 + TotalCountof2s * 2 + TotalCountof1s * 1)}{No.ofuser} \tag{6.5}$$

With respect to NDCG@10 evaluation metrics, we observe the performance of all the queries. Performance evaluation discussion of each query of all platforms are:

Query1:

With an average Google time-bound search result, the NDCG@10 value is 0.929 greater than a Twitter search (0.7263) and our platform (0.924). Albeit, on our platform the percentage of users who gave our system a rank of very good is still higher than those for a Google time-bound search and a Twitter search (Table 6.4). Similarly, the percentage of Twitter search results ranked very good is higher than Google time-bound search results. The reason behind these differences goes back to the nature of Normalized DCG (NDCG). In NDCG, the when highly relevant documents appear on top they always get higher weightage than lower results. In a Google time-bound search result, the percentage of participants who ranked the first result as high quality (very good) is higher than the other sources. Therefore, the value of Google here is higher than either of the other platforms.

Table 6.4: Evaluation matrix (nDCG) for a query 1

	Google Time-bound	**Twitter**	**SHS(Social Health Signals)**
DCG	9.12	9.68	12.72
IDCG	9.81	13.33	13.76
nDCG	0.929	0.726	0.924

Query2:

In the second query, the NDCG@10 values are: 0.786412612 for a Google time-bound search, 0.916328092 for a Twitter search, and 0.92936253 for our system. In this survey, participants responses for very good are almost equally spread out for both Googles time-bound search and our systems search results. However, the difference between both values is significant because a higher portion of our results are rated very good. Additionally, this

time Twitters search NDCG@10 value is more in line with our system because the first top result of the Twitter search ranked very good over both platforms. Please see following table (Table 6.5).

Table 6.5: Evaluation matrix (nDCG) for a query 2

	Google Time-bound	**Twitter**	**SHS(Social Health Signals)**
DCG	10.03	9.67	13.15
IDCG	12.76	10.55	14.15
nDCG	0.78	0.91	0.92

Query3:

In the third query, the NDCG@10 value of a Google time-bound search is 0.98838052 (near to 1.0) because the first results has been ranked high quality by number of participants and remaining results are ranked as good and bad. The values of our system (0.852668155) and a Twitter search are (0.847387026) almost equal, even though the percentage of people ranking our systems results very good is higher than for Twitter search (Table 6.6). The top results of our system ranked very good, as compared to Twitter search results, which also ranked very good, but appear at end of the top ten list.

Table 6.6: Evaluation matrix (nDCG) for a query 3

	Google Time-bound	**Twitter**	**SHS(Social Health Signals)**
DCG	10.76	10.75	11.47
IDCG	10.89	12.69	13.45
nDCG	0.98	0.84	0.85

Discussion and Future Work

In this chapter, we will discuss the several decisions that directed our research. We will discuss the impact of the decisions and alternatives in the Section 7.1. We whttps://www.sharelatex.com/projec discuss unsolved problems and uncovered topics in Section 7.2.

7.1 Discussion

Twitter is very popular for research on various topics including in a healthcare. We have used tweets and a URLs content (mentioned in the tweets) for finding health information. Users often share URLs in tweets when they wish to give more detailed information than the length limitations of tweets generally allow. We have observed that 70% of daily tweets (specifically, in the case of those related to diabetes) contain URLs. Similarly, a study [] shows that 30% of all tweets contain at least one URL. With our survey we observed that 80% of the participants ranked tweets which didnt have any links as very bad. This suggests that people have a preference for tweets with URLs when seeking information; with that notion in mind, we chose to further explore the content of URLs to see what additional details and trends we might uncover. To gauge the reliability of the information source, we initially looked to both author expertise (a user who posted message on Twitter) and to the URL domain. However, we eventually chose to use the URL domain as our judgement criteria/metric for reliability. To computationally approximate a Twitter users level of expertise in a health domain or in a given field would be its own area of research.

One of the key challenges for users of Twitter is judging the expertise of other users to select trustful information and credential sources about health topics. In our research, we observe that there are many popular URLs in the result set which have a significantly lower domain rank (0, 1, or 2 according to Google domain rank API); however, we chose to exclude them in favor of those with domain pageranks greater than four.

We used a heat-map to visualize tweet traffic through use of both the tweet origin (physical location from which a tweet is sent) and a users location (physical location as identified in the Twitter profile). On the heat-map, larger values were represented by small dark gray or black squares (pixels) and smaller values by lighter squares. After checking the number of diabetes cases of the US at the state level [10], we found that there is a correlation between the number of people tweeting about diabetes in a state and the number of reported diabetes cases.

7.2 Future Work

In this thesis, we extract relevant and reliable documents based on a user query in near real-time. We plan to extend this thesis further by including semantic categorisation in which the results will be placed into different groups (drug, medication, symptom, etc.) using UMLS MetaMap. UMLS MetaMap is a program developed at the National Library of Medicine (NLM) to map biomedical text to the UMLS Metathesaurus or concepts referred to in the text. We will incorporate the domain knowledge using UMLS-Metathesaurus (Unified Medical Language System) to annotate the search results. With this addition, users would be able to filter out results according to any desired categories. The techniques that will be used in the implementation of this system are principally based on domain semantics, knowledge bases (UMLS, WordNet) and Semantic Web techniques. For example, we used taxonomy based approach for (a) data collection, (b) search query understanding, and (c) data annotation and retrieval. We will also use ontological knowledge and

domain knowledge from UMLS to perform semantic categorization of health information into health categories. Additionally, one of the aims of the system is to create a domain-specific knowledge base to help better serve the OHIS. However, we have not evaluated our system for complex queries such as those questions which contain many words. A more complex query makes it difficult for the system to create a triple. We have proposed one approach to mitigate this issue, which is combining multiple words in a question for a triples token. For example, in the question How does obesity cause increase the mortality rate? we can combine the words cause and increase and use them as a predicate. Also, we can combine adjectives and nouns. We can also work to improve the performance of dynamic queries as these queries are executed with data from six-hour intervals, which is time consuming, by increasing clusters.

7.3 Conclusion

Twitter has changed the traditional way of sharing and seeking health information for healthcare professionals and general users. All kinds of information is available on a Twitter for OHIS. According to [11] this study, OHIS use two online services, a Web search engine and social media search engine, to find out information. When OHIS want facts and a deeper understanding of information, they look for a Web search engine. In the case of social media search engines (e.g., Twitter) OHIS are interested in real-time content, breaking news, trends, and information which contains a social perspective. There are many challenges in the existing microblog search services, and they are as follows:

The results are limited to keyword-based techniques to retrieve relevant health information for a given query; sometimes the results do not contain real-time information. They use only messages or posts as a data source to find information. Ranking of the results is based on relevancy. Does not consider reliability of the sources of information.

In our approach, we have addressed these issues to extract reliable, relevant health

information from Twitter in near real-time since there are challenges due to the real-time nature of a Twitter (velocity), information overload (volume), and noise in the textual data. In our thesis, we have addressed these problems by using state-of-the-art approaches such as semantics-based pattern mining, a similarity-based algorithm on query expansion, ranking of results by content popularity (social media share and like counts), and reliability (Google domain pagerank). To check the results relevancy, we looked at two similarity algorithms; the first is a bag-of-words model, and the second is a vector-based model. The bag-of-words model we have chosen is a Jaccard index algorithm; for our vector-based model, we have chosen a TF-IDF cosine similarity algorithm. After an evaluation (precision and recall), a TF-IDF cosine similarity with a Random Forest classifier performed best for our experiments. However, more queries need to be tested before we can assess the best model to obtain relevant information for OHIS. 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 assess the performance of SHS with respect to Twitter search. In all three studies, user preferred content from SHS over Twitter search. Our assumption that if a relevancy evaluation (survey) were performed, the percentage of participants who would rank our results very good would be higher than the results recorded for Twitters own search results is proven correct. Relevancy and popularity are the main criteria used to rank the results. Also, we have observed, Twitters search returns results in order of those which are deemed most likely to be relevant to a users query. A Google ranking is based on relevance and back links. The higher quality the links, the higher the Google pagerank. However, in our case, we selected social media share and like counts as our popularity measures. SHS a very comprehensive system with the motivation of aiding users in keeping track of health information. The system developed with contribution to public health systems as well as social media based systems.

Bibliography

[1] Apache. Apache lucene core, 2015. [https://lucene.apache.org/core/.

[2] Apache. HDFS Architecture Guide, 2015. http://hadoop.apache.org/docs/r1.2.1/hdfs_design.html.

[3] Apache. Lifecycle of a storm topology, 2015. http://storm.apache.org/documentation/Lifecycle-of-a-topology.html.

[4] Apache. Storm, distributed and fault-tolerant realtime computation, 2015. [Online; accessed 22-February-2015].

[5] Apache. Tutorial, 2015. [Online; accessed 22-February-2015].

[6] Brenner and Smith. 72% of online adults are social networking site user (pew research centers internet american life project rss), August 5, 2013.

[7] Simon Carter, Manos Tsagkias, and Wouter Weerkamp. Twitter hashtags: Joint translation and clustering. 2011.

[8] Pew Research Center. Health fact sheet, January, 2014.

[9] Cynthia Chew and Gunther Eysenbach. Pandemics in the age of twitter: content analysis of tweets during the 2009 h1n1 outbreak. *PloS one*, 5(11):e14118, 2010.

[10] Disease Control and Prevention. Diagonised diabetes, age adjusted rate., 2012. http://gis.cdc.gov/grasp/diabetes/DiabetesAtlas.html.

[11] Munmun De Choudhury, Meredith Ringel Morris, and Ryen W White. Seeking and sharing health information online: Comparing search engines and social media. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*, pages 1365–1376. ACM, 2014.

[12] Jadhav et al. Online information seeking for cardiovascular diseases: A case study from mayo clinic. *Studies in health technology and informatics*, 205:702, 2014.

[13] Fox and Purcell. Chronic disease and the internet, March 24, 2010.

[14] Brian Honigman. 24 outstanding statistics on how social media has impacted health care (physician referral management software referralmd), 2013-09-02. https://getreferralmd.com/2013/09/healthcare-social-media-statistics/.

[15] Zhuoye Ding Qi Zhang XuanJing Huang. Automatic hashtag recommendation for microblogs using topic-specific translation model. In *24th International Conference on Computational Linguistics*, page 265. Citeseer, 2012.

[16] IBM. Infosphere streams text analytics, 2015. http://www.ibm.com/developerworks/library/bd-streamstextanalytics/.

[17] Ashutosh Jadhav, Amit Sheth, and Jyotishman Pathak. Analysis of online information searching for cardiovascular diseases on a consumer health information portal. In *AMIA Annual Symposium Proceedings*, volume 2014, page 739. American Medical Informatics Association, 2014.

[18] Ashutosh Sopan Jadhav, Hemant Purohit, Pavan Kapanipathi, Pramod Anantharam, Ajith H Ranabahu, Vinh Nguyen, Pablo N Mendes, Alan Gary Smith, Michael

Cooney, and Amit P Sheth. Twitris 2.0: Semantically empowered system for understanding perceptions from social data. 2010.

[19] Cher Han Lau, YueFeng Li, and Dian Tjondronegoro. Microblog retrieval using topical features and query expansion. In *TREC*. Citeseer, 2011.

[20] Ilene MacDonald. Healthcare professionals flock to twitter, April 22, 2014.

[21] Matteo Magnani, Danilo Montesi, Gabriele Nunziante, and Luca Rossi. Conversation retrieval from twitter. In *Advances in Information Retrieval*, pages 780–783. Springer, 2011.

[22] Survey monkey. Rating and ranking average calculations, 2015. http://help.surveymonkey.com/articles/en_US/kb/What-is-the-Rating-Average-and-how-is-it-calculated.

[23] National Library of Medicine. Umls quick start guide. 2014. http://www.nlm.nih.gov/research/umls/quickstart.html.

[24] Ottenhoff. Infographic: Rising use of social and mobile in healthcare, December 17, 2012.

[25] Alexander Pretschner and Susan Gauch. Ontology based personalized search. In *Tools with Artificial Intelligence, 1999. Proceedings. 11th IEEE International Conference on*, pages 391–398. IEEE, 1999.

[26] Johnson N. Melton S. Shaffer-Hudkins, E. Social media use by individuals with diabetes, 2013.

[27] Amit Sheth, Ashutosh Jadhav, Pavan Kapanipathi, Chen Lu, Hemant Purohit, Gary Alan Smith, and Wenbo Wang. Twitris: A system for collective social intelligence. In *Encyclopedia of Social Network Analysis and Mining*, pages 2240–2253. Springer, 2014.

[28] RJP Stronkman. *Exploiting Twitter to fulfill information needs during incidents*. PhD thesis, TU Delft, Delft University of Technology, 2011.

[29] Jaime Teevan, Daniel Ramage, and Merredith Ringel Morris. # twittersearch: a comparison of microblog search and web search. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 35–44. ACM, 2011.

[30] TF-IDF. Jaccard index, 2015. http://www.tfidf.com/.

[31] Twitter. About twitter, 2015. https://about.twitter.com/company.

[32] Twitter. The streaming apis, 2015. https://dev.twitter.com/streaming/overview.

[33] Wikipedia. Apache hadoop, 2015. http://en.wikipedia.org/wiki/Apache_Hadoop.

[34] Wikipedia. Discounted cumulative gain, 2015. https://en.wikipedia.org/wiki/Discounted_cumulative_gain.

[35] Wikipedia. Jaccard index, 2015. http://en.wikipedia.org/wiki/Jaccard_index.

[36] Wikipedia. Machine learning, 2015. http://en.wikipedia.org/wiki/Machine_learning.

[37] Wikipedia. Random forest, 2015. https://en.wikipedia.org/wiki/Random_forest.

[38] Wikipedia. Tfidf, 2015. http://en.wikipedia.org/wiki/Tfidf.

[39] Wikipedia. Twitter, 2015. http://en.wikipedia.org/wiki/Twitter.

[40] WordNet. Wordnet, 2015. https://wordnet.princeton.edu/.