ISSN 2348 – 7968

Big Data Analysis – Design and Implementation of Novel Algorithms for Tweets Retrieval and processing

Prof. Sharath P C[1], Prof. Spoorthi K M[2]

[1,2]Assistant Professor, Department of CS&E, TJIT,
Bangalore, 560083, India.

Abstract

Twitter as a micro blogging platform has vast potential to become a collective source of intelligence that can be used to obtain opinions, ideas, facts, and sentiments. The issue on collective intelligence conversation retrieval with activated knowledge-base decision making is addressed. The methodology differs from the existing literature in the sense that analysis is on Twitter micro blog messages as opposed to traditional blog analysis in the literature which deals with the conventional blogosphere. Another key difference in the methodology is that the visualization techniques are applied in conjunction with artificial intelligence-based data mining methods to classify messages dealing with the trend topic. The methodology also analyzes demographics of the authors of such Twitter messages and attempt to map a Twitter trend into what's going on in the real world. The findings reveal a pattern behind trends on Twitter. The findings also enable to understand the underlying characteristics behind the 'trend setters', providing a new perspective on the contributors of a trend.

Keywords: *Twitter, Micro blogs, Trends, SNS, Tweets.*

1. Introduction

Twitter is a popular online social networking service where people often share information or opinions about personalities, politicians or products. Twitter as a micro blogging platform has vast potential to become a collective source of intelligence that can be used to obtain opinions, ideas, facts, and sentiments. Users post short text messages called tweets, which are limited by 140 characters in length and can be viewed by user's followers. One interesting thing about Twitter is that it has a section on its main page entitle Trending Topics, which displays the top mentioned terms on Twitter at any given moment. This is generated based on Twitters proprietary algorithm, but nonetheless provides an interesting zeitgeist into the events talked about by the Twitter community. When a new topic becomes popular on Twitter, it is listed as a trending topic, which may take the form of short phrases e.g. Michael Jackson or hashtags e.g. #election. What the Trend2 provides a regularly updated list of trending topics from Twitter. It is very interesting to know what topics are in trend and what people in other parts of the world are interested in. However, a very high percentage of trending topics are hash tags, a name of an individual, or words in other languages and it is often difficult to understand what the trending topics are about, it is therefore important to classify these topics into general categories for easier understanding of topics and better information retrieval.

The issue on collective conversation retrieval with activated knowledge-base decision making is addressed. The attempt to dissect the anatomy of a trending topic to find out what makes it tick, gather information about the posts mentioning the topic to a maximum of up to 1500 posts and from that, the data is analyzed to investigate any patterns that occur in trending topics. More specifically, text messages or posts are only basic bricks composing more complex and socially relevant conversations between communities of users. While it is certainly interesting to analyze each post on its own, it is also very important to be able to manipulate these complex structures, e.g. Clustering conversations, retrieving the conversations about a given topic, and understanding the topic of a conversation. Evidently, all these information retrieval capabilities are based on the viability of a model to rank a set of conversations with respect to some information requirements.

The social networks and micro blogs which are products of Web 2.0 are becoming the tool of choice for information dissemination, sharing and interpersonal communication and networking. It is a new platform being rapidly adopted by all walks of life, from politicians to businessmen, young and old, for use in citizen journalism to being a medium to stay close to friends and family. By harnessing this information from online social network and micro blogging sites such as Twitter, an understanding about the collective "wisdom of crowds" is obtained, and leverages its data in policymaking, decision support, economic analysis, epidemic behaviour and various other applications.

2. Literature Survey

The literature survey carried out reveals a fair amount of research done in the area of big data analysis. Micro blogging is a broadcast medium in the form of blogging. A micro blog differs from a traditional blog in that its content is typically smaller in both actual and aggregate file size. Micro blogs allow users to exchange small elements of content such as short sentences, individual images, or video links. These small messages are sometimes called micro posts [17]. Twitter is a online social networking service and micro blogging service that enables its users to send and read text-based messages of up to 140 characters, known as tweets [17]. A word, phrase or topic that is tagged at a greater rate than other tags is said to be a trending topic.Trending topics become popular either through a concerted effort by users or because of an event that prompts people to talk about one specific topic. These topics help Twitter and their users to understand what is happening in the world [17].

Lucene is a high-performance, full-featured, java, open-source, text search engine API written by Doug Cutting. Lucene is specifically an API, not an application. This means that all the hard parts have been done, but the easy programming has been left to users. The payoff for you is that, unlike normal search engine applications, you spend less time wading through tons of options and build a search application that is specifically suited to what user doing. You can easily develop a custom search application, perfectly suited to your needs. Lucene is startlingly easy to develop with and use [18].

JSON (JavaScript Object Notation) is a lightweight data-interchange format. It is easy for humans to read and write. It is easy for machines to parse and generate. It is based on a subset of the JavaScript Programming Language, Standard ECMA-262 3rd Edition - December 1999. JSON is a text format that is completely language independent but uses conventions that are familiar to programmers of the C-family of languages, including C, C++, C#, Java, JavaScript, Perl, Python, and many others. These properties make JSON an ideal data-interchange language. JSON is built on two structures: A collection of name/value pairs. In various languages, this is realized as an *object*, record, struct, dictionary, hash table, keyed list, or associative array. An ordered list of values. In most languages, this is realized as an *array*, vector, list, or sequence. These are universal data structures. Virtually all modern programming languages support them in one form or another. It makes sense that a data format that is interchangeable with programming languages also be based on these structures [19].

2.1 Accessing Tweets from the Twitter

Accessing Tweets from a Twitter is primary for building a database to get processed and extract information. Twitter has three types of API's REST API, Search API and Streaming API. REST API defines a set of functions to which the developers can perform requests and receive responses via HTTP protocol. Because it uses HTTP, REST can be used practically for any programming language and easy to test. The streaming API requires that you keep the connection active. This requires a server process with an infinite loop, to get the latest tweets. The search API is the easier of the two methods to implement but it is rate limited. Each request will return up to 100 tweets, and you can use a page parameter to request up to 15 pages, giving you a theoretical maximum of 1,500 tweets for a single query.

2.2 Conversation Modeling

On a very simple level of abstraction the on-line conversation is made of a series of messages exchanged between users using an on-line SNS. Given that level of abstraction conversations happening in SNSs may be described as very similar to many other off-line technological mediated forms of communication. On the contrary, SNS-based interactions have some unique features that must be considered while developing a model for conversations. In a SNS, space is defined by four properties persistence, replicability, search ability and addressed to an invisible audience.

These properties generate a set of specific dynamics in SNS interactions and create a specific background for conversational practices, with persistence of characteristics of a message being available online after the first publication for an undefined time. Persistence is strengthening by the search ability of on-line digital contents. Every contents published on-line is not only permanently accessible but also potentially easier and easier to find thanks to improving searching algorithms and techniques. In Figure.1 the main differences between off-line and online conversations related to these unique features is pointed out. Squares represent the first message of a conversation and diamonds follow-up messages [1].Synchrony is not required the time Δt between the original message becomes interested in a specific topic he/she will always be able to search on-line for that topic and even restart a conversation, which became inactive long time before. In the figure.1, the time elapsed between the first message and the first reply may be much longer in the on-line case Joining an on-line conversation allows the new user to have a complete view of what has been said until that time. In the figure.1 Bob joins the conversation on the message marked with a cross. In the off-line case, he will know only the messages exchanged from that point, while on an online conversation his scope will. Cover all the interactions occurred before his decision to join the conversation. On-line persistence of the whole conversation, including the original messages, can bring many users to post messages not directly addressed to the last comment available but referring to any previous message. This may end up in many concurrent conversations starting from a single message. This cannot be modeled as a quasi-chain structure with a rigid chronological sequence of interactions where almost every message refers to the previous one, as it usually happens with off-line conversations. Finally, the on-line environment enables the collaboration of many more people compared with typical off-line physical conversational environments.

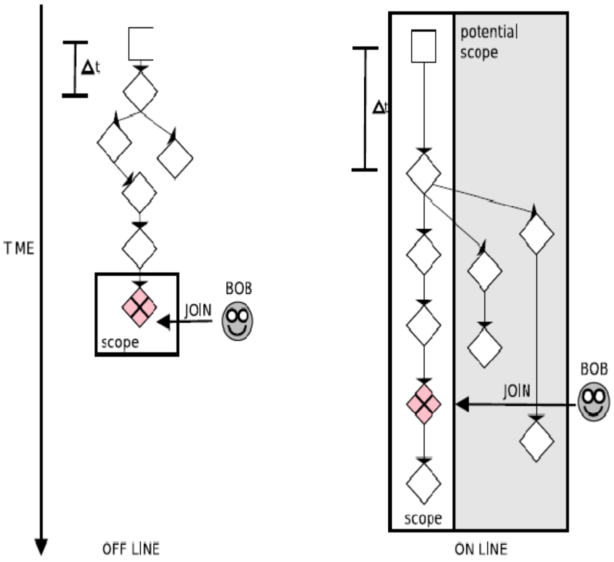

Figure.1 Main Differences Between On-Line And Off-Line Conversations

Having these features of on-line conversations can provide formal model describing them. The basic communication step of a conversation involves an actor performing a communicative act in the on-line environment at a precise timestamp. The focus on communicative acts expressed as a persistent communicative object that can later interpreted by a set of other actors to whom the object is available.

It follows that a polyadic conversation is a chronological sequence of text messages exchanged between actors where the people involved may change during the conversation. Each message will refer to a previous one, constituting a tree-structure [1].

2.3 Related Work

In [1], a novel search paradigm for micro blogging sites resulting from the intersection of Information Retrieval and Social Network Analysis (SNA). This approach based on a formal model of on-line conversations and a set of ranking measures including SNA centrality metrics, time-related conversational metrics and other specific features of current micro blogging sites. The ranking approach has been compared with other methods and tested on two well known social network sites (Twitter and Friend feed) showing that the inclusion of SNA metrics in the ranking function and the usage of a model of conversation can improve the results of search tasks.

The author formalizes the concept of conversation in the context of Social Network Sites. Traditional information retrieval technique can extend to deal with these social aspects [2].

The authors in [3], describes the problem of finding opinionated tweets about a given topics is retrieved. Where automatically construct

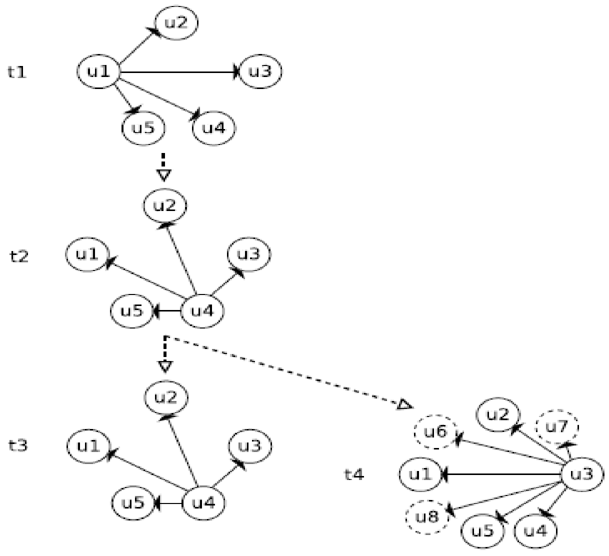

Figure.2 A Graphical Representation of A Polyadic Conversation

Opinionated lexica from sets of tweets matching specific patterns indicative of opinionated messages. When incorporated into learning to- rank approach, results show that this automatically opinionated information yields retrieval performance comparable with a manual method. Finally, topic-related specific structured tweet sets can help improve query-dependent opinion retrieval.

In [4], retrieving information from Twitter is always challenging given its volume, inconsistent writing and noise. Existing systems focus on term-based approach, but important topical features such as person, proper noun and events are often neglected, leading to less satisfactory results while searching information from tweets. This paper proposes a novelty feature extraction algorithm which targets the above problems, and presents the experiment results using TREC11 dataset. The proposed approach considers both term-based and pattern- based features and distributes weights accordingly. The paper experiment four different setting to evaluate different combinations and results show that the approach outperformed traditional method of using term-based or pattern only methods and signify the importance of topical features in micro blog retrieval.

In [5], context in a web-based social system can be a valuable source of user information. On Twitter, context can be derived from user interactions, content streams and friend- ship. The focus on extracting user context by means of conversation patterns and user-generated twitter lists. A novel approach, which utilizes just the context to extract twitter users topics of interest is presented. The approach achieved a precision of 84% indicating that user context can be exploited for topic information.

In [6], most Twitter search systems generally treat a tweet as a plain text when modeling relevance. However, a series of conventions allows users to tweet in structural ways using combination of different blocks of texts. These blocks include plain texts, hash tags, links, mentions, etc. Each block encodes a variety of communicative intent and sequence of these blocks captures changing discourse. Previous work shows that exploiting the structural information can improve the structured document e.g. web pages retrieval. A set of features, derived from the blocks of text and their combinations, is used into a learning-to-rank scenario. That show structuring tweets can achieve state-of-the-art performance. The approach does not rely upon social media features, but when do add this additional information, performance improves significantly.

In [7], twitter summarizes the great deal of messages posted by users in the form of trending topics that reflect the top conversations discussed. These trending topics tend to be connected to current affairs. Different happenings can give rise to the emergence of these trending topics. For instance, a sports event broadcasted on TV, or a viral meme introduced by a community of users. Detecting the type of origin can facilitate information filtering, enhance real-time data processing, and improve user experience. In the paper, authors have introduced a typology to categorize the triggers that leverage trending topics: news, current events, memes, and commemoratives. A set of straightforward language-independent features are formulated that rely on the social spread of the trends in order to discriminate among the types of trending topics. The method provides an efficient way to immediately and accurately categorize trending topics without need of external data, outperforming a content-based approach.

In [8], twitter provides a list of most popular topics people tweet about known as Trending Topics in real-time, it is often hard to understand what these trending topics are about where most of these trending topics are far away from the personal preferences of the twitter user. In this article, attention to the issue of personalizing the search for trending topics via enabling the twitter user to provide RSS feeds that include the personal preferences along with a twitter client that can filter personalized tweets and trending topics according to a sound algorithm for capturing the trending information. The algorithms used are the Latent Dirichlet allocation (LDA) along with the Levenshtein Distance. The experimentations show that the developed prototype for personalized trending topics (T3C) finds more interesting trending topics that match the Twitter user list of preferences than traditional techniques without RSS personalization.

In [9], user-contributed content is creating a surge on the Internet. A list of buzzing topics can effectively monitor the surge and lead people to their topics of interest. Yet a topic phrase alone, such as “SXSW”, can rarely present the information clearly. In this paper, a proposal to explore a variety of text sources for summarizing the Twitter topics, including the tweets, normalized tweets via a dedicated tweet normalization system, web contents linked from the tweets, as well as integration of different text sources. Concept-based optimization framework for topic summarization is employed, and conducts both automatic and human evaluation regarding the summary quality. Performance differences are observed for different input sources and types of topics. This also provides a comprehensive analysis regarding the task challenges.

In [10], although Twitter provides a list of most popular topics people tweet about known as Trending Topics in real time; it is often hard to understand what these trending topics are about. Therefore, it is important and necessary to classify these topics into general categories with high accuracy for better information retrieval. Twitter Trending Topics categorized into 18 general categories such as sports, politics, technology, etc. The experiment with 2 approaches for topic classification; (i) the well-known Bag-of-Words approach for text classification and (ii) network-based classification. In text-based classification method, the construct word vectors with trending topic definition and tweets, and the commonly used tf-idf weights are used to classify the topics using a Naive Bayes Multinomial classifier. In network-based classification method, identifying top 5 similar topics for a given topic based on the number of common influential users. The categories of the similar topics and the number of common influential users between the given topic and its similar topics are used to classify the given topic using a C5.0 decision tree learner.

In [11], user-contributed messages on social media sites such as Twitter have emerged as powerful, real-time means of information sharing on the Web. These short messages tend to reflect a variety of events in real time, making Twitter particularly well suited as a source of real-time event content. In this paper, approaches for analyzing the stream of Twitter messages to distinguish between messages about real-world events and non-event messages. The approach relies on a rich family of aggregate statistics of topically similar Message clusters. Large-scale experiments over millions of Twitter messages show the effectiveness of the approach for surfacing real-world event content on Twitter.

In [12], social networking media generate huge content streams, which leverage, both academia and developers efforts in providing unbiased, powerful indications of users’ opinion and interests. Here, presents Cloud4Trends, a framework for collecting and analyzing user generated content through micro blogging and blogging applications, both separately and jointly, focused on certain geographical areas, towards the identification of the most significant topics using trend analysis techniques. The cloud computing paradigm appears to offer a significant benefit in order to make such applications viable considering that the massive data sizes produced daily impose the need of a scalable and powerful infrastructure. Cloud4Trends constitutes an efficient Cloud-based approach in order to solve the online trend tracking problem based on Web 2.0 sources. A detailed system architecture model is also proposed, which is largely based on a set of service modules developed within the VENUS-C research project to facilitate the deployment of research applications on Cloud infrastructures.

In [13], users create short messages pertaining to a wide variety of topics. Certain topics are highlighted by Twitter as the most popular and are known as “trending topics.” In this paper, the outline methodologies of detecting and identifying trending topics from streaming data have been designed to collect the data from Twitter’s streaming API and put into documents of equal duration. Data collection procedures will allow for analysis over multiple time spans, including those not currently associated with Twitter-identified trending topics. Term frequency-inverse document frequency analysis and relative normalized term frequency analysis are performed on the documents to identify the trending topics. Relative normalized term frequency analysis identifies unigrams, bigrams, and trigrams as trending topics, while term frequency-inverse document frequency analysis identifies unigrams as trending topics.

The authors in [14], that twitter has received significant research interest lately as a means for understanding, monitoring, and even predicting real-world phenomena. However, most existing work does not address the sampling bias, simply applying machine learning and data mining algorithms without an understanding of the Twitter user population. In this paper, a first look at the user population themselves, and examined the population along the axes of geography, gender, and race/ethnicity. Overall, found that Twitter users significantly overrepresented the densely population regions of the United States, are predominantly male, and represent a highly non-random sample of the overall race/ethnicity distribution. Going forward, the study sets the foundation for future work upon Twitter data. Existing approaches could immediately use the analysis to improve predictions or measurements. By enabling post-hoc corrections, the work is a first step towards turning Twitter into a tool that can make inferences about the population as a whole. More nuanced analyses on the biases in the Twitter population will enhance the ability for Twitter to be used as a sophisticated inference tool.

The authors have discussed in [15], the dynamic squarified treemap for visually representing the trending topics on Twitter. The main ingredients for this graph are the speed of tweets and the acceleration of them being published and thus have developed algorithms to calculate both of them. Moreover, a simple clustering algorithm to deal with grouping related topics in online twitter streams. The final representation in a dynamic squarified treemap fills the gaps the dynamic squarified treemap forms a powerful visual tool to visualize trending topics based on the list provided by Twitter. However, trending topics. The analysis in this paper has been done on the currently working on a system in which monitors a sample of the twitter stream and detect trending topics. The system calculates the speed and acceleration every second and updates the screen accordingly. Based on the size and rate of growth of a cluster of words / topics the dynamic squarified treemap serves as an early warning system for trends.

In [16], a methodology for visualizing text streams in real time. The approach automatically groups similar messages into “countries,” with keyword summaries, using semantic analysis, graph clustering and map generation techniques. It handles the need for visual stability across time by dynamic graph layout and Procrustes projection techniques, enhanced with a novel stable component packing algorithm. The result provides a continuous, succinct view of evolving topics of interest. It can be used in passive mode for overviews and situational awareness, or as an interactive data exploration tool. To make these ideas concrete, describing their application to an online service called Twitter Scope.

3. Description of Project Work

The tweets retrieved from twitter website is observed as trending topics based on popularity gained. On particular trending topics, gather information about the posts mentioning the topic to a maximum of up to 1500 posts, and retrieve the conversation chains when tweets respond to others, and a new approach of analyzing the trending retrieved tweets form twitter is made based on the decision making stratagem.

3.1 Problem Definition

An attempt to retrieve of conversations from twitter website based on the trending topics, where the data can be very large. The topics should be retrieved from the ongoing conversation in the real world from the twitter on random.

The attempt to analyze the structured tweets of a trending topic to find out what makes it 'tick' selecting a specific topic. Specifically, random topics are selected which appear in the top category of the trending topics list, gather information about the tweets or posts mentioning the topic to a maximum of 1500 posts or more posts, and from there the data is analyzed to investigate any patterns that occur in trending topics. The data of trending topics can be Retweeted, Replied for the tweets, Trended tweets.

3.2 Hardware Requirements

PROCESSOR: PENTIUM IV 2.4GHZ
HARDDISK: 80 GB
RAM: 512 MB

3.3 Software Requirements

OPERATING SYSTEM: Windows XP, Linux
FRONT END: Eclipse
Lucene
MySQL JDBC Driver
Tomcat
Java SDK
MySQL DBMS
Twitter for Java

PROGRAMMING LANGUAGE: JAVA\J2EE, SQL

4. Design Methodology

4.1 System Design (High Level Design)

High-level design provides an overview of an entire system, identifying all its elements at some level of abstraction. Such an overview is important in a multi-project development to make sure that each supporting component design will be compatible with its neighbouring designs and with the big picture. A high-level design document will usually include a high-level architecture diagram depicting the components, interfaces and networks that need to be further specified or developed.

4.2 Architecture

A Conversation Retrieval system is made of four applications: One Web/Application Server (interface between users and system).
One Conversation Server (storage, indexing and search of conversations).
One Trend Server (gets trending topics and distributes them among tweet Retrieval clients).
One or more Tweet Retrieval Clients (to retrieve tweets in parallel).
For an efficient and effective installation every application should run on a dedicated computer, and several Tweet Retrieval Clients should be executed in parallel. However all servers and clients may be installed on a single computer for testing.
The following figure illustrates the system architecture. The Trend Server repeatedly gets from the Twitter API the current trending topics. These topics are distributed to the Tweet Retrieval Clients that use the Twitter Search API to get the corresponding tweets, and also to retrieve the conversation chains when tweets respond to others. These are sent to the Conversation Server that stores the tweets on a relational database and indexes them. Users may then query the system through a Web application.

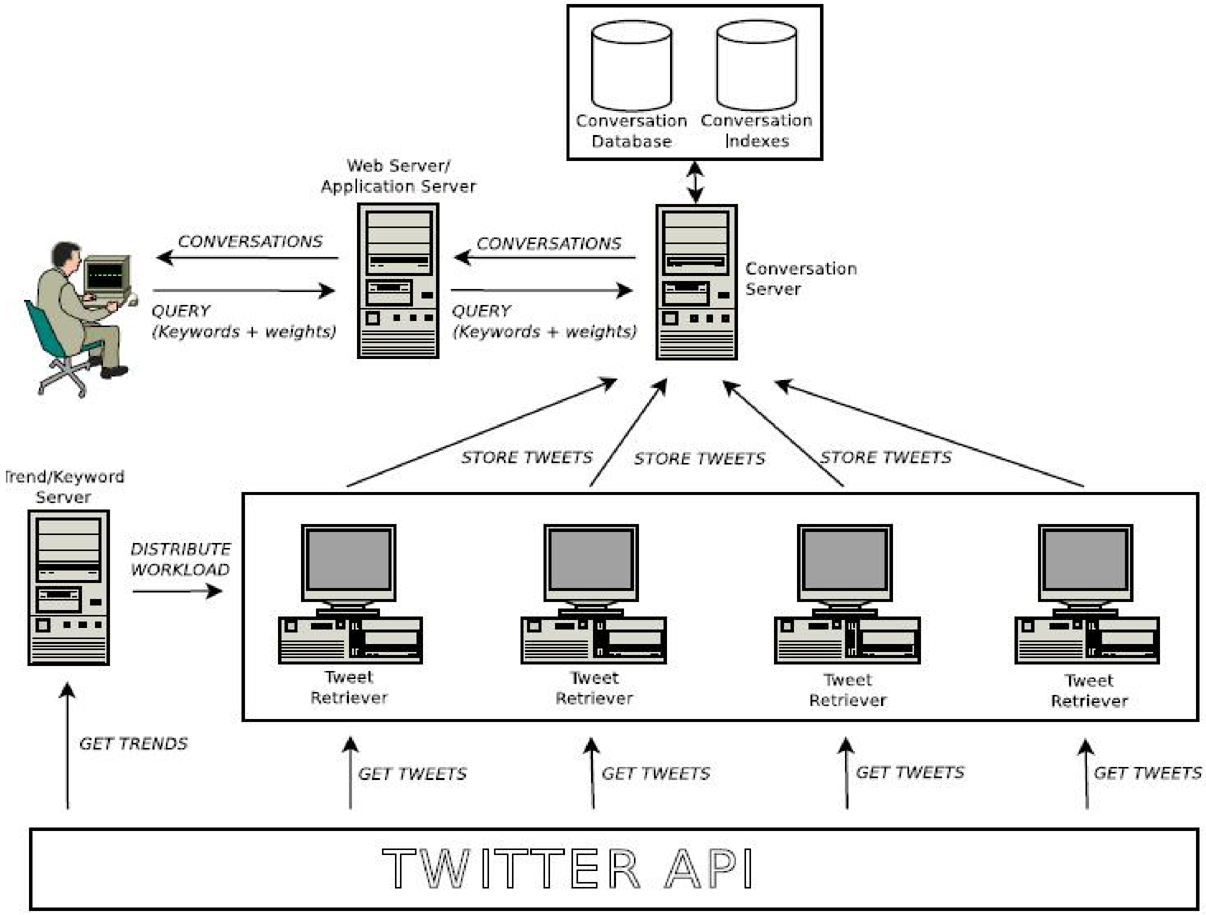

Figure 3: System Architecture of Conversation Retrieval for Twitter

4.3 Detailed Design (Low Level Design)

Low Level Design (LLD) is like detailing the High Level Design(HLD). It defines the actual logic for each and every component of the system. Class diagrams with all the methods and relation between classes comes under LLD. LLD describes each and every module in an elaborate manner.

4.4 Implementation Plan

Module 1: Conversation Server

Create a database named conversation. Using the conversation database create two tables user and tweet table. The Conversation Server consists in three main java servers:

CRStoreServer: used for storing of retrieved tweets.

CRSearchServer: used for searching tweets.

CRConversationServer: used for retrieving tweets.

Module 2: Trend Server

A file named 'KEYWORD' should be created and write line by line the keywords you want to monitor what are called trending topics. Launch the trend server.

Module 3: Tweet Retriever Client

Using JSON and Twitter4j and with respective ip address and port number assigned tweet retriever client should be launched. The client should have started retrieving and indexing tweets and can check if this is happening by looking at the tweet table in the MySQL server.

5. Testing

Testing is a process of executing a program with the intent of finding an error. It presents an interesting anomaly for the software engineering. It is a set of activities that can be planned and conducted systematically. Software testing is often referred as 'Verification and Validation'.

5.1 Types of Testing

Unit Testing: In Unit Testing, each module will be tested individually and integrate with the overall system. Unit testing focuses verification efforts on the smallest unit of software design in the module. This is also known as module testing. The module of the system is tested separately. This testing is carried out during programming stage itself. In this testing step each module is found to be working satisfactorily as regard to the expected output from the module. There are some validation assist in identifying all errors and bugs. The sample data are used for testing. checks for fields also. It is very easy to find error debut in the system.

System Testing: Testing of the debugging programs is one of the most critical aspects of the computer programming, without that the system would never produce the output for which it was designed. Testing is best performed when the user development are asked to assist in identifying all errors and bugs. The sample data are used for testing.

Validation Testing: At the culmination of the black box testing, software is completely assembled as a package, interfacing error have been uncovered and corrected in final series of software tests. Validation testing can be defined many ways, but a simple definition is that validation succeeds when the software functions in the way that can be reasonably expected by the customer. Validation succeeds when the software functions in the way that can be reasonably expected by the customer.

TABLE 1
TEST CASES

Test ID	Test case Description	Expected output	Observed Output
1	CREATION OF DATABASE	DATABASE SHOULD BE CREATED WITH TWEET AND USER TABLE	SAME AS EXPECTED OUTPUT
2	CONFIGURATION OF CONVERSATION SERVER	CONVERSATION SERVER AND TWITTER API COMMUNICATION SHOULD BE SUCCESSFUL	SAME AS EXPECTED OUTPUT
3	CONFIGURATION OF TREND SERVER	USING KEYWORDS SHOULD COMMUNICATE WITH TWITTER API	SAME AS EXPECTED OUTPUT
4	CONFIGURATION OF TWEET RETRIEVAL CLIENTS	BASED ON TRENDS TWEETS SHOULD START DOWNLOADING WITH TWEETED USER INFORMATION	SAME AS EXPECTED OUTPUT

FUTURE WORK

Implementation of a web server that is interface between the user and the conversation system, which is an advantage to access the tweets directly through the webpage and give the trend name directly in the search tab. Easy to handle and performance can be improved.

CONCLUSION

A new approach is introduced with some preliminary aspects of what is called as conversation retrieval, which is an information retrieval activity that exploits structural aspects in addition to the exchanged text messages. It is an approach of analyzing Trend patterns for the Twitter micro-blogging platform. The analysis done on the trending topics based on the list of keywords provided and analysis done on the retrieved tweets and twitter users from the twitter API. The approach exploited the information retrieval on the collective intelligence characterized in the Twitter message pool and user base, and applied decision-making stratagem of the set of Twitter users contributing towards the discussion of a particular trend. This is potentially useful in the areas of policymaking, decision support, economic analysis, business intelligence, marketing, epidemic research and their related fields.

REFERENCES

[1] Matteo Magnani, Danilo Montesi, Luca Rossi, *Conversation retrieval for microblogging sites*, Springer Link, 2012.

[2] Matteo Magnani, Danilo Montesi, *Toward conversation retrieval,* Italian research conference on digital library management systems, 2010.

[3] Zhunchen Luo, Miles Osborne, Tingwang, *Opinion retrieval in twitter*, Association for the Advancement of Artificial Intelligence, 2012.

[4] Cher Han Lau, Yuefeng Li, Dian Tjondronegoro, *Microblog retrieval using topical features and query expansion,* Queensland University of Technology, 2011.

[5] Ravali Pochampally,Vasudeva Varma, *User context as a source of topic retrieval in twitter*, SIGIR 2011 Workshop on Enriching Information Retrieval, 2011.

[6] Zhunchen Luoy, Miles Osbornez, Sasa Petrovicz and Tingwangy, *Improving twitter retrieval by exploiting structural information*, Association for the Advancement of Artificial Intelligence (www.aaai.org), 2012.

[7] Zubiaga, Damianospina, Víctorfresno, Raquel Martínez, *Classifying trending topics: a typology of conversation triggers on twitter*, ACM, 2011.

[8] Jinan Fiaidhi, Sabah Mohammed, *Towards identifying personalized twitter trending topics using the twitter client rss feeds*, Journal of emerging technologies in web intelligence, Vol. 4, no. 3, 2012.

[9] Fei Liu, Yang Liu, Fuliangweng, *Why is "sxsw" trending? Exploring multiple text sources for twitter topic summarization*, Proceedings of the Workshop on Language in Social Media, 2011.

[10] Kathy Lee, Diana Palsetia, Ramanathan Narayanan, MD. Mostofa Ali Patwary, Ankit Agrawal, and Alok Choudhary, *Twitter trending topic classification*, 11th IEEE International Conference on Data Mining Workshops, 2011.

[11] Hila Becker, Mor Naaman, Luis Gravano, *Beyond trending topics: real-world event identification on twitter*, Association for the Advancement of Artificial Intelligence (www.aaai.org), 2011.

[12] Vakali, Maria Giatsoglou, Stefanos Antaris, *Social networking trends and dynamics detection via a cloud-based framework design*, Athena MSND'12 Workshop, 2012.

[13] James Benhardus, *Streaming trend detection in twitter*, UCCS REU for artificial intelligence, 2010.

[14] Alan Mislove, Sune Lehmann, Yong-yeol Ahn, Jukka-pekka Onnela, J. Niels rosenquist, *Understanding the demographics of twitter users*, Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media, 2011.

[15] Sandjai Bhulai, Peter Kampstra, Lidewij Kooiman, Ger Koole, Marijn Deurloo and Bert Kok, *Trend visualization on twitter: what's hot and what's not?*, DATA ANALYTICS: The First International Conference on Data Analytics, 2012.

[16] Emden R. Gansner, Yifanhu, and Stephen North, *Visualizing streaming text data with dynamic maps*, IEEE VGTC, 2012.

[17] http://en.wikipedia.org/wiki

[18] http://jakarta.apache.org/lucene

[19] http://www.json.org