ABSTRACT
Twitter is the most popular microblogging site. It provides us with real-time data. This article provides a survey of techniques for retrieving information from the Twitter stream. These techniques aim to find relevant and real-world information with respect to the query. For retrieving the most relevant information, used query expansion techniques. Twitter data contain a large amount of information. Information rank retrieval techniques find important data and give the final score to that information with respect to user interest profiles.

Keywords
Real-time data, relevance information, microblog, Twitter stream.

1. INTRODUCTION
1.1 Introduction
Microblog is a broadcast medium that allows users to post short and frequent messages [8]. It's a new communication medium compared with traditional data, microblogging has gained increased attention among users, organizations, research scholars in different disciplines.

Twitter is currently the fastest-growing microblogging service, with more than 140 or 150 million users producing over 400 or 500 million tweets per day [8]. It's an unfiltered feed of status or tweets, no more than 140 characters to a network of followers using various communication services. Tweets size are limited; Twitter is updated millions of times a day by users all over the world [8], and its data varies hugely based on user interest and behaviors. So Twitter data have a large amount of information scaling from the latest news, events etc.

Twitter provides timely information of any event. Observing, keeping, and analyzing this content of user-generated data can yield new unprecedented important information, which is not available from traditional media [8]. Tweets do the live reporting of events [13] means finding the information what people are talking away from some conferences, debates, sporting events etc.

This Article provides a survey of retrieving information from microblog using various techniques and gives the most relevance score.

1.2 Challenges
Twitter has reported everything from daily life stories to real-world events. Millions of tweets are updated, so people have no time to visualize all those tweets.

A major problem is there is no restriction to post a tweet, update information or status so many people provide false, incorrect information about some events. Large number of spelling and grammar errors, and the use of not a proper sentence structure and mixed language so people can't distinguish important data from unused data. Not all tweets are relevant to the user query or interest profile.

One-way communication. Twitter often acts as a one-way communication platform. Twitter is used by celebrities, TV shows, companies and websites to simply get the word out. It is not used for relationship-building.

2. MICROBLOG REAL TIME FILTERING
A user has interest in real-time topics, events, and they want to stay up to date in that topic using the stream of microblog posts [2].

Main goal of real-time filtering is monitoring the stream of microblog with respect to user interest profile (query) and finding relevance score.

Provide interesting content to user by [1, 2]
1. Push notification on a mobile phone: Interest content might be shown to the user through mobile phone notification.
2. Periodic email digest: User interest profile might be aggregated into an email form that periodically sent to a user [1]. In that case a user could read a longer story about content.

2.1 System Overview
In this section introduce system architecture for retrieving tweets and do the scoring of tweets based on query.
System contain four component [2].

I. Feature Extraction Component: It extract features from twitter based on TREC-API (Stream API and Rest API). After obtaining twitter stream we apply preprocessing and filtering to reduce tweets we need to process.

II. Feature Representation Component: It represent and expand semantic feature by different expansion techniques.

After extracting tweet we need to represent those feature in proper format so it is suitable to calculate relevance score between tweet and profile.

III. Candidate Generation Component: We classify tweets into the most relevant profile or remove it directly if it does not match any profile.

IV. Scoring and pushing component: By the semantic feature (consider verbs and nouns in tweet text) and social media attributes we got two score semantic (Ci) and quality (Qi) so final score Si = CiQi.

3. QUERY EXPANSION

The query provided by the user is not in a structured and that is incomplete. So then we need to expand that query and do the correct for the better relevance information.

Main problem in retrieval is that query is short and unable to accurately describe user’s information needs. So solution of this problem is query Expansion [4, 9].

For exemplar, rather than writing “earth quake”, people may instead use the word “quake” or directly include a hashtag such as “#eq” in message [4]. It is not a realistic to manually recognize the whole set of related keyword and categorization tag (i.e., hashtags) for each user profile.

Simple methods of query expansion,

- Finding exactly or nearly the same as another word of query words, and discovering for the synonym as well
- Finding all the various morphological forms of words be caused by each word in the search query[10]
• Fixing grammatical mistake, free from spelling error and automatically finding for the corrected form or forward for consideration it in the results
• Re-weighting the terms in the original query[10]

4. RELEVANCE MODEL
The topic represent as a triple of a title that contains few keywords, description that summarizes the topic in one sentences and a narrative that consist of a paragraph that gives the more details. Based on this we retrieve relative or relevance data.

4.1 Tf-Idf Score
Which calculates the similarity between a tweet and a profile in vector space model with TFIDF weight of terms [1]. Tweets and profiles can be expressed as vectors.

TFIDF is the product of two statistics[17]

• Term Frequency
• Inverse document frequency

4.1.1 Term Frequency (TF)
TF measure the number of times a term occurs in a document. But all documents are different in length so on a large document the frequency of the terms will be much higher compare to small document. So we require to normalize the document based on its size[17].

For that divide the TF by total number of terms.

Term frequency of word i in document j is

\[ t_{ij} = f_{ij} / N_{tj} \]

Here \( f_{ij} \) = Frequency of term i in document j.

\( N_{tj} = \) Total Term in Doc 1

4.1.2 Inverse document frequency
In TF all term are consider equally important. But such a term like the, is, that etc. is not important and may seems lots of times so we require a way to weigh down the effect of more time appearing or occurring terms [17]. Also the words that happen less in the documents can be closer.

So Inverse document frequency is

\[ idf_i = 1 + \log_2 (N / df_i) \]

Here \( df_i \) = document frequency of term i

\( N \): total number of documents

The tf-idf weight of a particular word is the yield of its tf weight and its idf weight.

\[ W_i = t_{ij} \cdot idf_i = t_{ij} \cdot \log_2 (N / df_i) \]

Score of a document with respect to a query:

\[ \text{Score}(q,d) = \sum_{i=1}^{n} W_i \]

4.2 BM25
BM25 is a bag of words ranking function[16] that ranks an information based on the user interest profile or query words appearing in each document information[1,2].

Developed in the context of the Okapi system in London University.

BM25 formula contains many parameter which need to be tuned from relevance assessment [15].

Given a user interest profile \( P \), containing keywords \( p_1, p_2, p_3, ..., p_n \) the BM25 score of a document \( D \) is

\[ \text{Score}(D, P) = \sum_{i=1}^{n} \frac{\text{IDF}(p_i)}{(f(p_i,D) + k_1) \cdot (1 + b \cdot \frac{|D|}{avgdl})} \]

Where \( f(P, D) \) is \( p_i \)’s term frequency in the document \( D \), \( |D| \) is the length of the document \( D \) in words, and \( \text{avgdl} \) is the average document length in the text collection from which documents are drawn[19]. \( k_1 \) and \( b \) are default parameters, usually chosen, in absence of an advanced optimization, as \( k_1 \in [1.2, 2.0] \) and \( b \in [0.5, 0.8] \)[19]. IDF (qi) is the IDF weight of the query term qi[19].

4.3 Language Model
Main Goal Of language Model is assign a probability of a sentence or sequence of words [18]. If sequence of word is \( w_1, w_2, w_3, ..., w_n \) then

\[ P(W) = P(w_1, w_2, w_3, ..., w_n) \]

We can also find probability of upcoming word based on previous word

\[ P(w_{n+1} | w_1, w_2, w_3, ..., w_n) \]

Here, \( P(W) \) and \( P(w_{n+1} | w_1, w_2, w_3, ..., w_n) \) is language model.

Rank each documents by the probability of particular document given a query [18].

\[ P(D|Q) = P(Q|D) \cdot P(D) / P(Q) \]

(Here \( P(Q|D) = \bigwedge P(q_i|D) \) and \( P(q_i|D) = q_i \cdot D|D] \)

5. CONCLUSION
In this paper survey the research in the area of information retrieval on microblog. Its aims finding most relevance information respect to user query and provide real-word occurrence. As a fast growing microblog site, Twitter provides us to valuable and real content of any real event. Using query expansion technique we can retrieve better result. This article display major challenges of twitter and describe different technology which used in different paper for expand query and then retrieve tweet. And give the relevance score.

6. FUTURE WORK
The research can be continued in use a language model and word2vec both for query expansion and providing more training to the system so we can get more accurate result and provide highly relevance tweet to the user.

7. REFERENCES


[9] Cher Han Lau, YueFeng Li, Dian Tjondronegoro, Microblog retrieval using topical features & query expansion, Queensland University of Technology

[10] Ashish Kankaria, Query Expansion techniques, Indian Institute of Technology Bombay, Mumbai


[12] Jimmy Lin, 1 Miles Efron 2, Yulu Wang, 3, Garrick Sherman 2, and Ellen Voorhees, Over view of the TREC-2015 Microblog Track, 1 University of Waterloo, 2 University of Illinois, Urbana-Champaign, 3 University of Maryland, College Park, 4 NIST


[14] Iadh- Ounis 1, Craig Macdonald 1, Jimmy Lin 2, 3, Ian Soboroff 4 * Over view of the TREC-2011 Microblog Track, 1 University of Glasgow, Glasgow, UK 2 Twitter, San Francisco, CA, USA 2 University of Maryland, College Park, MD, USA 4 NIST, Gaithersburg, MD, USA.


[16] Ben He and Iadh Ounis, Term Frequency Normalisation Tuning for BM25 and DFR Models, Department of Computing Science United Kingdom.


[18] Book of Information Retrieval by Christopher D.Manning, Prabhakar Raghavan and Hinrich Schütze