ORIENTAL JOURNAL OF
COMPUTER SCIENCE & TECHNOLOGY
An International Open Free Access, Peer Reviewed Research Journal
Published By: **Oriental Scientific Publishing Co., India.**
www.computerscijournal.org

ISSN: 0974-6471
December 2016,
Vol. 9, No. (3):
Pgs. 165-176

Discovering Spatio-temporal Patterns of Themes in Social Media

TOBORE IGBE, BOLANLE OJOKOH* and OLUMIDE ADEWALE

Department of Computer Science, Federal University of Technology, P. M. B. 704, Akure, Nigeria.
*Corresponding author E-mail: bolanleojokoh@yahoo.com

http://dx.doi.org/10.13005/ojcst/09.03.02

(Received: December 14, 2016; Accepted: December 17, 2016)

ABSTRACT

Social networking website creates new ways for engaging people belonging to different communities, moral and social values to communicate and share valuable knowledge, therefore creating a large amount of data. The importance of mining social media cannot be over emphasized, due to significant information that are revealed which can be applied in different areas. In this paper, a systematic approach for traversing the content of weblog, considering location and time (spatiotemporal) is proposed. The proposed model is capable of searching for subjects in social media using Boyer Moore Horspool (BMH) algorithm with respect to location and time. BMH is an efficient string searching algorithm, where the search is done in such a way that every character in the text needs not to be checked and some characters can be skipped without missing the subject occurrence. Semantic analysis was carried out on the subject by computing the mean occurrence of the subject with the corresponding predicate and object from the total occurrence of the subject. Experiments were carried out on two datasets: the first category was crawled from twitter website from September to October 2014 and the second category was obtained from spinn3r dataset made available through the International AAAI (Association for the Advancement of Artificial Intelligence) Conference on Web and Social Media (ICWSM). The results obtained from tracking some subjects such as Islam and Obama shows that the mean occurrence of the analysis of the subject successfully reveals the pattern of the subject over a period of time for a specific location. Evaluation of the system which is based on performance and functionality reveals that the model performs better than some baseline models. The proposed model is capable of revealing spatiotemporal pattern for a subject, and can be applied in any area where spatiotemporal factor is to be considered.

Keywords: Boyer-Moore-Horspool Alogrithm, Search processing, Spatio temporal pattern, Sementic analysis.

INTRODUCTION

Social Networks have become very popular in recent years because of the increasing proliferation and affordability of internet enabled devices such as personal computers, mobile devices and other more recent hardware innovations such as internet tablets. Social network sites vary greatly in their features and user base. Beyond users and friends, some have photo-sharing or video-

sharing capabilities; others have built-in blogging and instant messaging technology. These sites create new ways for engaging people belonging to different communities. Some example of social media includes: Facebook, Tweeter, LinkedIn, Instagram, YouTube, Myspace, Foursquare and so on. They also provide a very powerful medium for communication among individuals that leads to mutual learning and sharing of valuable knowledge (Baumeret al., 2010).

Social Media provides a wealth of social network data. For example, social networks may contain links to posts, blogs or other news articles, which can be mined in order to discover useful business information and applications (Charu, 2011). Themes in social media refer to majorfocus or topic of discussions among people in social media, which can either be a person or people, event, crises, disease, and so on. Tracking themes in social media will reveal trends about topics and also expose compliment and technical issues faced by consumers of products and services in a specific place for a particular period of time. Mining the content of social media can be utilized for search result summarization and the result can be fed into decision support system for use by web analysts, public opinion miners, and commercial organisations.

Theme tracking has become imperative due to its various advantages; it will help companies to know the experience of users' with their products and services, and also to get information that they can use effectively to stand out from competitors. This kind of information could be market trends, industry research, sales promotion, competitor analysis, and medical research among others (Mike & Steve, 2008). Spatiotemporal mining of themes in social media requires processing unstructured information contained in social media sites and extracting meaningful information based on location and time. Social media provides the platform for different people to create, share and exchange information and ideas from different communities, faith, ideology and at different times. This makes mining social media much more interesting, to help reveal important information based on trend and changes on a particular theme at different times for different communities (Sowjanya *et al.,* 2014).

Existing researches have considered some areas in weblog analysis, in terms of textual analysis (Zielinski et al., 2013; Kumar *et al.*, 2004) with regards to spatiotemporal consideration(Sowjanya *et al*., 2014); some have worked on sentiment analysis (Ojokoh *et al*., 2012; Kumar *et al*., 2004; Pang & Lee, 2008; Mike et al., 2009; Jayanta & Abhisek, 2011), topic modelling and busty event detection; none have reported theme analysis in weblog. The occurrence of a theme with respect to predicate and object shows how much the theme is prominent for time duration in a particular location. The result from this analysis can be useful in political campaign to answer questions such as to what political party or politician is getting the most attention of bloggers. It also shows what a particular device from a manufacturer does which can be used for personalised marketing. This paper describes a method for discovering spatiotemporal patterns of tracked themes from tweeter. Our approach can be applied to any social media with text content, and have feature that keep track location and time of its social content. Most social media maintain the time and either directly or relatively the geographical location where social content was added (posted) (Roick and Heuser, 2013). Boyer Moore Horspool (BMH) algorithm is adapted to track the presence of theme in social media; and semantic analysis is carried out to discover the presence of pattern for a particular location over a period.

The content of this paper is structured as follows: In section two,we make a review of works focusing on structural analysis and sentiment analysis of text contentfrom social media. Our proposedmethod is described in section three. Sectionfour discusses the use of Boyer Moore Horspool algorithm for tracking themes in weblogs. The experimental results and evaluation are discussed in section five, while conclusion and recommendations for further works are presented in section six.

Related works

Online social media contains a large collection of data from many viewpoints, ranging from politics, to entertainment, sports,and so on. We focus our literature review on structural,semantic analysis and spatio-temporal analysis of social media text content. Structural analysis refers

to exploring thetext contentin social media to discover hidden information, while semantic analysis investigates and extracts intelligence from social media. Spatio-temporal analysis discusses techniques used in location and time analysis of social media data.This research work entails both structural analysis (searching for theme) from social media text and semantic analysis, finding the pattern of the theme with respect to location and time.

Social Media Structural Analysis

Social media structural analysis involves revealing useful information from social media data by examining the characteristics and pattern of the content. The analysis considers either or both the text and the network created in social media to uncover information hidden in social media.

Jayanta and Abhisek (2011) addressed the problems of identifying themes in Social Media and detecting sentiments using statistical text-mining of blog entries. The crux of the analysis lies in mining quantitative information from textual entries. Once the relevant blog entries for a company or its competitors are filtered out, theme identification is performed using highly accurate novel techniques termed as 'Best Separator Algorithm'. Logistic regression coupled with dimension reduction technique (singular value decomposition) was used to identify the tonality of those blogs.

Combinational approach has also been employed for text mining and analysis with appreciable amount of success. For instance, in Zielinski *et al*., (2013), combined techniques of Trustworthiness Analysis (TA), Multilingual Tweet Classification (MTC) and Geo-parsing (GEO) were used in social media text mining and network analysis for decision support in natural crisis management. The result of this research indicatedsuccessin tracking themes in social media and revealing spatiotemporal pattern of the themes.

Hassan *et al*. (2009) proposed a method for event detection and tracking in social streams. Events and stories are characterized by a set of descriptive, collected keywords. Intuitively, documents describing the same event will contain similar sets of keywords, and the graph of keywords for a document collection contains clusters of individual events. A network of keywords was built based on their co-occurrence in documents. Furthermore, new event detection algorithm which creates a keyword graph and uses community detection methods analogous to those used for social network analysis to discover and describe events was developed. The work confirmed that the constellations of keywords describing an event may be used to find related articles.

The identification of popular and important topics discussed in social networks is crucial for a better understanding of societal concerns. Budak *et al*. (2011), considered network topology for trends detection to be able to distinguish viral topics from topicsthat are diffused mostly through the news media. Their approach was based on structural concept called coordinated and uncoordinated trends that uses friendship information to identify topics that are discussed among clustered and distributed users respectively. Their technique recorded in a gain in efficiency and was within an acceptable error bound.

Although, the performance and resultsof these works were impressive but the absence of location and time features which is an important factor in weblog was not considered. This factor would have provided anexcellent explanation to the results obtained.

Social Media Sentiment Analysis

Sentiment analysis, also called opinion mining, is the analysis of people's opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes (Liu, 2012).

A method to extract emotions from social media was developed by Mike *et al*. (2009). In his work, gender differences of participant in social networks were studied. A classification scheme was constructed to quantify the extent to which positive and negative emotions were expressed in each comment. In particular, "I miss you" could be interpreted as positive and almost a synonym of "I love you", even though it may suggest sadness.

Similarly, "I love you" or "love you" is ostensibly a very strong positive emotion but seems to be used relatively casually in MySpace. As a result of issues like these, a set of classification guidelines was constructed in the form of a list of phrases and associated suggested classifications. However, there were no consideration for aggregating of emotions based on regions of participant and the time these emotional comment was expressed to determine how emotions are expressed in different locations and over a time period.

Another work related to ours is Pang& Lee (2008); they presented traditional automatic sentiment detection techniques in social media. They used lexicon-based methods that rely on a collection of known and precompiled sentiment terms. In addition, they considered unsupervised lexicon approach that was capable of revealing sentiment based on degree of positivity of a text using subjective indicator. Their result was able to express emotions about a particular area to an individual.However, these relatively successful techniques often fail when moved to new location, because they are inflexible regarding the ambiguity of sentiment terms.

Ojokoh *et al*. (2012) tackled the problem of reading overwhelming large number of reviews available online about product to be purchased by customers and provide a quick description and summary of the performance of the product. This makes it possible for customers to make better and quick decision, and also help manufacturers improve their products and services. Each customer review is disintegrated into simple sentence segments using punctuation separator. The segmented reviews are analysed and categorised based on the feature set found in the database. Part of speech tagging is then carried out on the segmented reviews, and then polarisation is carried out based on the opinion word, to classify it as either positive or negative. Feature segments aggregation is then performed on the polarised review segments to determine final recommendation decision about the product. However, the reviews considered were not categorised into location, which would have assisted consumers make better decision about a product based on where they reside.

Spatio-temporalAnalysis

This analysis focuses on retrieving useful hidden information in social media through technique based on the principle of dimensionality reduction of documentto a new feature space in which the features are typically a linear combination of the features in the original data for further processing and analysis based on location and time.

Bayesian generative model called Location Aware Topic Model (LATM) which incorporate location was presented by Wang et al. (2007). The model is inference is achieved by variational expectation-maximization (EM) for location aware topic. A fixed number of location labels is used in place of real latitudes and longitudes, and they assume that each term is associated with a location label. A topic assignment is first generated for each word in a document, with respect to a multinomial distribution. Then the term and the location are generated dependent on this topic assignment, according to two different multinomial distributions.

Mei *et al*. (2006) propose a Spatiotemporal Theme Pattern model, where every word is either taken from a general background topic or from a time and location related language model. This model is centred on Probabilistic Latent Semantic Indexing (PLSA). Inference is performed via expectation-maximization (EM). Moreover, the mixture coefficients between the background topic and other spatio-temporal topics ones is tuned manually. There is no prior is assumption of the distribution of the topic, since the model uses PLSA. Evaluation is carried out by showing anecdotal results.

Latent Dirichlet Allocation (LDA) modelling technique presented by Blei *et al*. (2003) is applied in text corpora for discovering latent semantic topics in large collections of text documents. The key insight into LDA is the premise that words contain strong semantic information about the document. Therefore, documents on roughly similar topics will use the same group of words. This modelling technique is highly modular and can also be applied in audio and video documents. Thus, the

LDA method can also be used in order to model the topic distribution of a new document more robustly, even if it is not present in the original data set. Despite the similarity of this model with PLSA in task performances, clustering, categorization and retrieval, the result from this model have comparative advantage over PLSA.

Williams *et al*. (2013)addressed Latent Periodic Topic Analysis (LPTA) in social network community or cluster, which tackled problems in constructing networks from unstructured data, analysing the community structure of a network, and inferring information from networks. The model considered two categories for detecting change of topic: first is the entropy of the movement between clusters, and second is the entropy of movements within clusters, each of which is weighted respectively by the frequency with which it occurs in the particular partitioning. Through the use of Graph analytics, it was able to achieve promising results on real-world data.

Topics-Over-Time (TOT) pattern was the focus of Wang and McCallum (2006) research model. They presented an LDA-style topic model that captures not only the low-dimensional structure of data, but also how the structure changes over time. Unlike other recent work that relies on Markov assumptions or discretization of time, here each topic is associated with a continuous distribution over timestamps, and for each generated document, the mixture distribution over topics is influenced by both word co-occurrences and the document's timestamp. Thus, the meaning of a particular topic can be relied upon as constant, but the topics' occurrence and correlations change significantly over time. They presented results on nine months of personal email, 17 years of Neural Information Processing Systems (NIPS) research papers and over 200 years of presidential state-of-the-union addresses, which showed improved topics, better timestamp prediction, and interpretable trends.

Our research on discovering spatio-temporal pattern of themes in social media takes inspiration from a number of sources. It is most similar to the work of Jayanta and Abhisek (2011) in terms of tracking subjects from copious blogs, comment and article generated by users of social networking sites. However, a different technique termed Boyer Moore Horspool (BMH) algorithm is employed to track subjects, taking into consideration regional and temporal factors in searching for subjects, which was not considered in Jayanta and Abhisek (2011) and in some reviewed literatures. The analysis of the tracked subjects will reveal the pattern of how these subjects disperse among different communities, which can be useful in many domains such as web analysis, business intelligence and public opinion monitoring.

METHODS

The operation of our model is divided into three sections: content pre-processing, search processing and result analysis. It is capable of searching for theme from the blog (web log of users' comment, post and news article) crawled social media sites and analysing the content, after removing unwanted elements such as tags from blog, where the theme is found to reveal the pattern for the theme occurrence. The theme to be searched for could be a person, a thing, place or idea capable of performing an action. Themes can be grouped into categories and some can occur in more than one category. Themes such as athletics, tennis, football can be classified as sports; while Ebola, nurse, virus, AIDS can be classified under

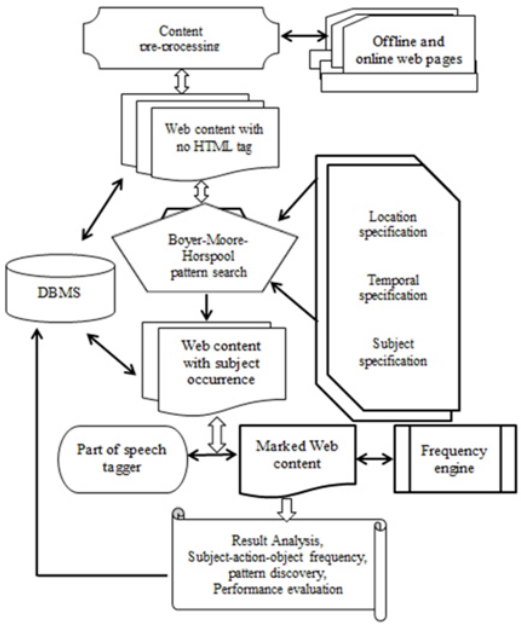

Fig.1: Proposed system architecture

health category. Communities are inevitably created in social media as a result of blog and sequences of replies to blog posted for different locations about a theme, which reveals the relationship that exist in that community. Also, similar theme tends to have different views over time as a result of circumstances and events unfolding at that time. (Wang and McCallum (2006); Wang *et al*. (2007)). The number of times the theme appears with respect to the action performed on an object can be referred to as the frequency of occurrence of the theme. The aggregate of the frequency over a period of time for a particular location shows the pattern for the theme. Significant information is revealed from social media when locations (communities) and time is taken into considerations (Mei *et al*. (2006);Williams *et al*. (2013)). Figure 1 depicts the proposed architecture of the system. The components of the architecture are interdependent of each other and their function is aimed at achieving the objective of the system.

Content Preprocessing

Web content with no HTML (Hyper-text Markup Language) tag constitutes the blog which represents text space to search for themes and they are obtained from social media website. These extracted texts are obtained from pre-processing offline or online web pages to remove unwanted web content such as images, graphics, table and cascading style sheet design, and so on. The website containing the blog can either be offline (crawling the content from an online site and saving it into a text file on the local system) or online (prepossessing the social media web pages directly from the World Wide Web location). Further processing is carried out to disintegrate the extracted blog into simple sentences for theme-searching operation.

Search Processing

Boyer-more-horspool pattern searching is the engine responsible for performing the search for themes from social media comments extracted from the pre-processing phase. It accepts three input parameters: TT=Temporal requirements, which could be one or more time periods (either months or years), and SS=region requirements, which could be one or more locations which forms the search criteria. These criteria impose restriction on the output of the search engine to satisfy both the SS and TT requirements. The third parameter is E=theme requirement; the search engine tracksE from sentences obtained from users' blog, the time and location features are spotted from the dataset using the demarcation parameter provided in the dataset. The outputs of the search engine are the filtered sentences where Eis located for the particular locations and time, given as:

$$H_{S,T}\,[match]=\{(E,SS_1,TT_1),(E,SS_2,TT_2) ,\ldots(E,SS_n,TT_n)\} \quad ...(1)$$

Where location and time are not specified, the output is clustered into locations and time where the theme is found. The obtained from the search operation is further labelled with different parts of speech using part of speech (POS) tagger to be able to distinguish subject, object and action in the sentence. In the analysis, the theme is the subject in the sentence which performs an action on an object.

Result Analysis

The frequency engine produces the performance analysis and result summary of the analysis and the analysis of the searched theme from the search processing phase. The result analysis shows the occurrence of the theme with corresponding object and the performed action. The output also shows the frequency of occurrence of the theme for a specific location and over a period of time.

Boyer Moore Horspool Algorithm

In this research, we adopt Boyer–Moore-Horspool (BMH) algorithm. Boyer and Moore (1977); Hume and Sunday (1991) affirmed that BMH is an efficient string searching algorithm that is the standard benchmark for practical string search literature. The search approach of the algorithm is done in such a way that every character in the text need not to be checked, some characters can be skipped without missing the theme occurrence (El-Mabrouk and Crochemore, 1996; Nebel, 2011). The algorithm pre-processes the theme being searched for (the pattern), P, but not the string being searched in (the text from social media),T. BMH algorithm is used to track themes from the comments in social media with respect to location and temporal features of these comments.

The algorithm uses information gained by pre-processing P to skip as many characters as possible during the search process, and it is defined as:

n=length(P) ...(2)

T[]=Break([.],T) ...(3)

m=length(T[x]) ...(4)

n and m are the length of characters in P and T[x] respectively, and P and T are the theme and extracted social media text respectively. T[] is an array of simple sentences obtained from T by breaking or splitting T using the full stop (.) delimiter, which marks the end of a sentence in English language. T[x] refers to a sentence at index x in T[].Equation (2) – (4) are used for initial processing to obtain parameters such as length of theme, and sentences from blog and the length of each sentence.

P[t] refers to the character at index t of string P, indexing begins at 0 to (n-1). T[i...j] refers to the substring of string T starting at index i and ending at j, inclusive. Equation (5) creates a lookup table generated by pre-processing P to determine the number of spaces to skip for each character in P where a miss-match occur, which isused by BMH model.Equation (6) performs the search process based on location and equation (7) groups the observed text where the match occurred into time segments.

$$G = \bigcup_{t=0}^{n} D(P[t]) \Rightarrow \begin{cases} n-1-t(n-1-t) > 1 \\ 1 \qquad otherwise \end{cases} \quad ...(5)$$

G is table lookup containing character-number pair, D is the function that maps () the occurrence of each of the character in P to the number of spaces to skip for mismatch during the pattern search operation. The table contains only unique character occurrences in P, and where there are multiple similar characters, only the last occurrence is kept in the table. Characters that are not present in P are added to the table and assigned n as the number of spaces to skip.

$$E[match]_s = \sum_{j=0}^{m} M_s\{G|k\}[p[0] \dots p[n-1] \rightarrow t[j] \dots t[j+k \dots m-1]] \quad ...(6)$$

The search process starts with an initial alignment where j is 0 of in P to in T at index k in T such that the last character of P is aligned with the character at index k of T. Subsequent search j will take position value from the lookup table to determine the number of characters to skip to align P in T. is the match occurrences of P in T in a particular location s which are values from 1,2,3, …, z representing different locations. The location and time factor are extracted as parameter within the same context as the text being processed. The location is available as longitude and latitude values, which is translated into normal text location using Twitter4J. The value of k is equal to n and it serves as an offset to determine where to align P in T. is a function that performs the tracking of P in T for a specified location using a table lookup.

$$H_{s,T}[match] = \bigcup_{s=1}^{SS} \bigcup_{t=1}^{TT} \varphi(t) E[match]_s \quad ...(7)$$

Where $\varphi(t) \in$ {day,month,year} is defined as a temporal coefficient, SS and TT are the spatial and temporal values respectively where the matches were observed, is the aggregation of the text (comment) containing matched theme taking into consideration space (location) and time.

A semantic theme analysis is performed by constructing sequences of subject-action-object triplets for matched themes after using part-of-speech (POS) tagging, which is the process of assigning part of speech to every word in a sentence. Part of speech includes: noun, verb, adverb, adjective, pronoun, conjunction and their sub-categories.

$HS,T[match]=W=w_1, w_2, \dots, w_y$ ...(8)

$XT=xt_1, xt_2, \dots, xt_x$ ...(9)

$P(XT \mid W)=p(xt_1, w_1), p(xt_2, w_2), \dots, p(xt_x, w_y)$ ...(10)

Where W is sentence in a social media, XT are sequence of tags, is a function that assigns tags in XT that are assigned to each word in W. Equation (8) – (10) assigns part of speech tag to the extracted sentence from social media where the theme was found. Equation (11) and (12) performs semantic analysis of the tracked theme which is the

subjectin the sentence with respect to the object and action (predicate) performed on the object by the subject.

$$F[P|W_a|W_o] = \sum_{i=1}^{K} (P, W_a, W_o \rightarrow H_{S,T}[match])_i \quad ...(11)$$

$F[P|W_a | W_o]$ is the frequency of stalked theme with corresponding actionand object found in () , is the action associated with the stalked theme, and is the object of the action. K is the total number of occurrences of the tracked subject. $P \subset H_{S,T}$ [match], $W_a \subset H_{S,T}$ [match] and $W_o \subset H_{S,T}$ [match].

The mean for the occurrence of the subject with corresponding predicate and object, can be obtained using the following expression:

$$M_p[W_a|W_o] = \frac{F[P|W_a|W_o]}{\sum H_{S,T}[match]} \quad ...(12)$$

Table 1: Summary of Extracted Dataset from Tweeter Website

Category	No of file	tweet extracted	File Name
Politics	14	4819	politics1 to politics14
Religion	11	4050	religion1 to religion11
Health	12	5858	health1 to health12

where $\sum H_{S,T}[\text{match}]$ is the total number of matches found in T.

EXPERIMENTS

We employ the proposed spatiotemporal theme tracking model to dataset that was obtained (crawled) from tweeter website using twitter4j, a Java application programming interface (API) that is designed to crawl content from tweeter website. The result obtained reveals that the proposed model performs satisfactorily for different type of theme in any category.

Dataset Construction

Twitter4J is a private Java library for the Tweeter API, which can easily be integrated into any Java application with the Tweeter service or as a standalone application (Twitter4J, 2014). Twitter4J API consists of features capable of crawling the content of the tweeter website for tweets on a specific category. In addition to specifying category, there is the option to specify time duration and the format of the returned content. There are two basic formats for the content returned by the tweeter4J API: JavaScript object notation format (JSON) or as plain text. The summary of the extracted data from tweeter website is displayed in table 1. The dataset consist of three folders representing the category of the dataset. Each folder contains files about the

Table 2: Result from tracking "Islam" theme in social media (religion category)

Week	Date	Location	Month	Dcount	$M_p[W_a\|W_o]$	Action	Object	$F[P\|W_a\|W_o]$
1	7th	UK	Sep-14	43	0.19	Is	Christen	8
		[93]			0.09	do	Member	4
2	14th				0.07	see	Member	3
					0.16	is	Muslim	7
					0.07	is	Cult	3
3	21th				0.09	see	People	4
4	30th				0.21	is	Religion	9
					0.12	can	Religion	5
5	5th		Oct-14	50	0.18	is	Religion	9
6	12th				0.2	chosen	Religion	10
					0.14	embrace	Muslim	7
7	19th				0.04	kills	Nonbelievers	2
					0.14	is	Muslim	7
					0.2	see	People	10
8	31th				0.1	make	Peace	5

information extracted from the site, and the size of the files ranges from 400 to 800kilobytes. The list of files in the same category contains the comment (tweets) and responses (retweets) to the related topic. Each file also contains descriptive information such as tweeter ID, screen name, date, location (place), and tweet url (uniform resource locator). The location is obtained from the longitude and latitude returned during the crawling process. The content of the files were extracted from September to October 2014.

Major events and news that could influence the content of the dataset crawled include: (1) the outbreak of Ebola epidemic disease that affected some West African countries and gradually entered United States of America (USA) and Europe. The epidemic was a challenge to the health system of affected countries and international health organisations such as World Health Organisation (WHO); (2) The increasing activity of political parties and politicians in Nigeria towards the general elections in February, 2015; and (3) efforts by USA and other developed countries to tackle the issue of terrorism, unemployment, food scarcity and global warming.

RESULTS AND DISCUSSIONS

Table 2 shows a section of information obtained from the extraction and analysis of tracking "Islam" theme from religion category. The table consists of nine (9) columns, the first column shows the week category and date refers to the date at the end of each week. The "Location" column indicates the location where the theme was found for the specified Date defined in the "Date" column and the total number of occurrence of the theme, enclosed in square bracket. The "Dcount" column shows the total number of occurrence of the theme for the specified month. The "$F[P|W_a \mid W_o]$" column gives occurrences for the specific "Action" performed on an "Object" by the theme. The "$M_p [W_a|W_o]$"column is obtained by calculating the mean from "$F[P|W_a|W_o]$"column for a particular time.

From the result obtained in table 2, we can tell that there was more occurrence of "Islam" in October than in September. Also, we can tell that there were more blogs discussing about Islam with relation to Christians and Muslims in September and religion and people in October.

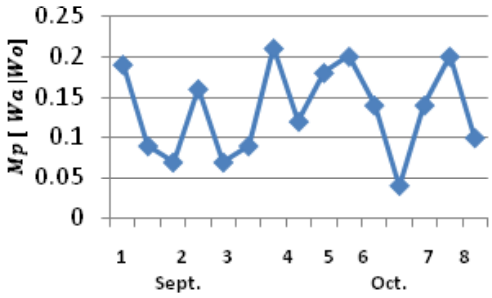

Fig.2: Mean occurrences for tracking "Islam" subject in UK

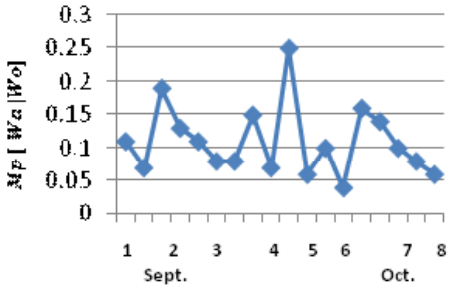

Fig. 3: Mean occurrences for tracking "Islam" subject in USA

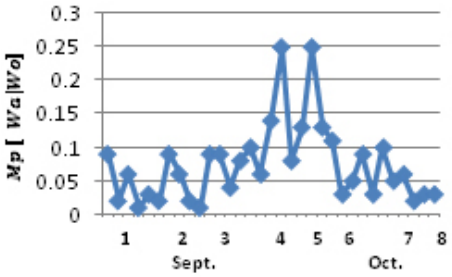

Fig. 4: Mean occurrences for tracking "Ebola" subject in USA

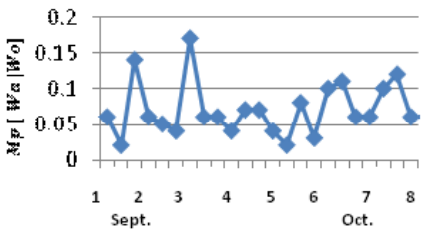

Fig.5: Mean occurrences for tracking "Ebola" subject in Liberia

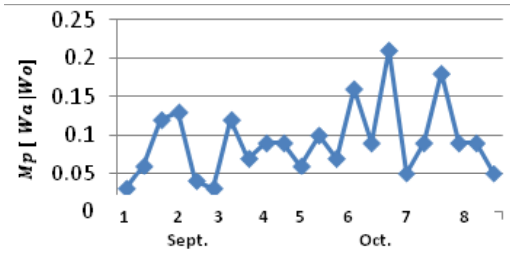

Fig.6: Mean occurrences for tracking "Ebola" subject in Nigeria

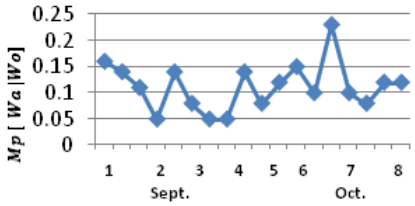

Fig.7:Mean occurrences for tracking "Ebola" subject in UK

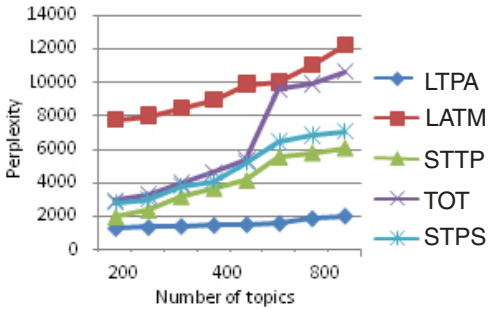

Fig.8: Perplexity comparison with baseline models

Figures 2 and 3 show the pattern for the occurrence of Islam in UK and USA respectively. There was a constant change in the discussion about Islam in UK but it dropped about the sixth to seventh week in October, compared to USA; there was a fairly stable change and there was an obvious occurrence between the fourth and fifth week.

Figures 4 - 7 show the pattern for the occurrence of ebola (a pandemic disease) in USA, Liberia, Nigeria and UK respectively. We noticed that the occurrence of the theme was more stable with close range values in Liberia than in any other place; this could be as a result of the persistence of the disease. In US, there was a peak occurrence between the fourth and sixth week. The peak occurrence in Nigeria can be seen between the sixth and eighth week (October), which was about the time the first case was diagnosed. Similarly, in USA the occurrence reached its climax at about the fourth and sixth week, this could be as a result of the possibility of the disease getting into the country.

Evaluation

The evaluation of our system (STPS) is based on measurable baseline of metric of perplexity. A brief about the baseline are as follows: Location Aware Topic Model (LATM) (Wang *et al*., 2007), Latent Periodic Topic Analysis (LPTA) (Williams *et al*., 2013), Latent Dirichlet Allocation (LDA) (Blei *et al*., 2013), Spatiotemporal Theme Pattern model (STTP) (Mei *et al*., 2006) and Topics-Over-Time model (TOT) (Wang and McCallum. 2006). These metrics measured the generalisation performance of a model, the ability of a model to generalise to unseen data.These metric measures the generalisation performance of a model, that is, the ability of a model to generalise to unseen data. The results from the baselines are adopted from Jiang and Hung (2013), and a corresponding setup of dataset is used in our research,the lower the perplexity measure, the better the generalisation performance. We set up an eight validation pre-set data to compare the models. Figure 6 shows the average perplexity of each model when the topic amount is set to different values. From the graph it is observed that when the topics is 500, the perplexity of STPS is 1502, while TOT, STTP, LATM and LPTA are 9885, 4180, 5398 ,5264 respectively.

CONCLUSION

Online social media contains weblogs that cover every facet of human endeavor.These are grouped into categories such as politics, education, entertainment, and so on. These categories have themes with spatiotemporal patterns; discovering these patterns are beneficial in many applications and domains such as weblog analysis, and public opinion mining. In this paper, we define the general problem of mining content of social media and propose a subject-based model for tracking theme and semantic analysis considering spatiotemporal patterns. The model uses Boyer Moore Horspool

algorithm to track themes from weblog, extract the text where the theme is found and perform semantic analysis on the text based on location and time.The evaluation result indicates the effectiveness of our model. Our approach can be used as an essential functionality for higher level tasks of analysis and research, based on spatiotemporal factors, such as prediction of user behaviour, reaction and activities. The different and emerging activities on social media makes it possible to discover more interesting details. There are still possibilities to extend this research; the direction to extend our work, will be to model miningtheme with spatio-temporal relationship between authors. This will reveal the connection between authors of comment in social media from different communities on the basis of analysing a tracked theme within a period of time. A theme can occur in more than one category, a spatio-temporal analysis of the same theme in different categories is another fascinating area of consideration.

REFERENCES

1. Baumer, E. P. S., Sinclair, j., & Tomlinson, B. (2010).Human factor in computingsystems. America is like metamucil: fostering critical and creative thinking about metaphor in political blogs. Atlanta, GA, USA, 34-45.
2. Boyer, R. S., & Moore, J. S., A fast string searching algorithm. **20**(10), 762-772, (1977). *Communications of Association for Computing Machinery(ACM)*, New York City.
3. Charu, C. A., Text mining in social networks in social network data analytics. (2nd ed.). *Springer*, 353-374, (2011).
4. Blei, D., Ng, A.,& Jordan, M.,Latent Dirichlet allocation, *Journal of Machine Learning Research*. **3**(1), 993–1022, (2003).
5. Budak, C., Agrawal, D., & El Abbadi, A., Structural trend analysis for online social networks. *Proceedings of the VLDB Endowment,* **4**(10), 646-656, (2011).
6. El-Mabrouk, N., &Crochemore, M. (Ed). (1996) Boyer-More strategy to efficient approximate string matching. Combinatorial Pattern Matching, Labuna Beach, California, France
7. Hassan, S., Hurst, M., & Alexey, M. (2009). Event Detection and Tracking in Social Streams. Proceeding of International AAAI Conference on Weblogs and Social Media.Third International AAAI Conference on Weblogs and Social Media. Retrievedfromhttp://aaai.org/ocs/index.php/ICWSM/**09**/paper/view/170/493
8. Hume A. and Sunday D., Fast String Searching software—Practice and experience, *ACM Digital Library.*, **21**(11), 1221–1248, (1991).
9. Jayanta, K.P., &Abhisek, S., Identifying themes in social media and detectingsentiments. I*nternational journal of statistics and applications*: **1**(1) 14-19, (2011).
10. Kumar, R., Novak, P. R., & Tomkins A.,Structure and evolution of blogspace. Commun.*ACM*,**47**(12):35-39, 2004,.
11. Leetaru, K. H., Culturomics 2.0: Forecasting Large-Scale Human Behavior Using Global News Media Tone In Time And Space. *Journal on the Internet*, **16** (9), (2011). Retrieved from:http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/3663/3040.
12. Liu, B., Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies,* **5**(1), 1-167, (2012).
13. Mei, Q., Liu, C., Su, H., & Zhai,C.X.(2006). A probabilistic approach to spatiotemporal theme pattern mining on weblogs, WWW.
14. Mike, K., & Steve, M., (2008). Centre for Business Performance. The use of information in decision making-Literature review for the audit commission.Cranfield, U.S.A.
15. Mike, T., David, W., &Sukhvinder, U. (2009). Data Mining Emotion in Social Network Communication: Gender differences in MySpaceStatisticalCybermetrics Research Group. School of Computing and Information Technology, University of Wolverhampton, Wulfruna Street, Wolverhampton WV1 1SB, UK.

16. Nebel, M. E. (2011). Search Texts-But Fast! The Boyer-More-HorspoolAlgorithm. *Springer-* VerlagBerlin Heidelberg Germany, 10.1007/978-3-642-15328-0_6.
17. Ojokoh, B. A., Olayemi, O. C., &Adewale, O. S., .Generating Recommendation Status of Electronic Products from Online Reviews, **4**, 1-10, (2012).*Intelligent Control and Automation*, doi:10.4236/ica.2013.41001.
18. Pang, B., & Lee, L., Opinion mining and sentiment analysis, **2**(1), 1-35, (2008). *Foundations and Trends in Information Retrieval*, U.S.A.
19. Roick, O., &Heuser, S., Location based social networks–definition, *current state of the art and research Agenda.* Transactions in GIS, **17**(5), 763-784, (2013).
20. Sowjanya, M., Ravindra, K., Kumar,R.Y.,(2014).Application of Concept-Based Mining Model in Text Clustering. *International Journal of Computer Science and Information Technologies*.**5**(5)6578-6582,(2014).
21. Twitter4J,(2014). Java library for the Twitter API2014.Retrieved from http://twitter4j.org/en/index.html.
22. Wang, X. & McCallum, A. (2006).Topics over time: a non-markov continuous-time model of topical trends, SIGKDD.
23. Wang, C., Wang, J., Xie X., and Ma W. Y. (2007). Mining geographic knowledge using location aware topic model.Proceedings of the 4th ACM Workshop On *Geographic Information Retrieval*, GIR. 65-70. DOI: 10.1145/1316948.1316967.
24. William, M. C, Charlie K. D., & Clifford J. W. (2013).Social Network Analysis with content and Graphs.
25. Zielinski, A., Middleton, S. E.,Tokarchuk, L. N., & Wang, X. (2013). Information systems for crisis response and management. Social media text mining and network analysis for decision support in natural crisis management.