|| Volume 3 || Issue 5 || May 2018 || ISSN (Online) 2456-0774

INTERNATIONAL JOURNAL OF ADVANCE SCIENTIFIC RESEARCH AND ENGINEERING TRENDS

Hot Event Evolution in Social Media by Understanding Short Text

Ms. Kavita K. Cholke [1], Mr. R. S. Bhosale [2]

Department of Information Technology, AVCOE, Sangamner[1]

Department of Information Technology , AVCOE, Sangamner Pune University , India[2]

Abstract— **Understand natural language texts is the need for machines highlights explosion information short text refer to text with limited context. Many applications just like web search and micro blogging services etc. having need to handle large amount of short text data. Understanding short text data bring tremendous value in social media mostly. Existing system having the challenges short texts do not always observe the syntax of written language texts. As a result traditional natural language processing tools, ranging from part- of-speech tagging to dependency parsing can not be easily applied. Second statistical signals are not sufficient for text mining. Short text is more ambiguous and noisy, and are generated in an enormous volume. Hence short text handling becomes complicated. I think semantic knowledge is required to understand short text in better way. In this paper system use semantic knowledge which is provided by a well known knowledgebase and automatically harvested from a web corpus. Proposed system overcomes the challenges by using text segmentation, part of speech tagging and concept labeling. our knowledge intensive approaches towards of both hot event evolution and discovering semantics of short text effectively and efficiently.**

Keywords: *Short text understanding, text segmentation, type detection, concept labelling, semantic knowledge, Micro blogging, Event Evolution, User Topic etc.*

I INTRODUCTION

Existing system having some problems to detect and evolution of event in social media. event evolution having problems of low search efficiency and low accuracy of results. There is difficult to tracking importance event which including identifying the influential spreaders. In this paper We have focus on short texts. We have propose context semantics when conducting text segmentation, and also hot event evolution model to identify the user interest. many application use short text for communication to each other such as microbloging and web search provide these services for people. Understanding of short text in better way will bring efficiency. Most important task of short text understanding is to discover hidden semantics texts. Named entity recognition locates text and classifies them into predefined categories such as persons, organizations, locations etc. "Latent topics" represent probabilistic distributions on words from a text . Knowledgebase are used to expressed "explicit topics" such as probabilistic distributions. However, categories, "latent topics" and "explicit topics "still having semantic gap with humans mental world. As stated in Psychologist Gregory Murphy's highly acclaimed book. Concept holdes our mental world together. Hence we have consider short text to detect concepts which is mentioned in a short text.

Short text understanding consist of three steps

1. **Text segmentation-**divide a short text into a collection of terms (i.e., words and phrases) contained in a vocabulary(e.g., "book Disneyland hotel California" is segmented as fbook Disneyland hotel California);

2.**Type detection-** determine the types of terms and recognize instances (e.g., both "Disneyland" and "california"are recognized as instances in Fig. 1, while "book" is recognized as a verb and "hotel" a concept);

3. **Concept labeling-** infer the concept of each instance (e.g.,"disneyland" and "california" refer to the concept theme park and state respectively in Fig.1

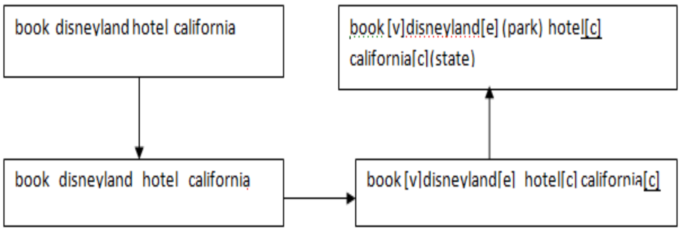
book disneyland hotel california
book [v]disneyland[e] (park) hotel[c]
california[c](state)
book disneyland hotel california
book [v]disneyland[e] hotel[c] california[c]

Figure.1 Example of short text Understand

The rapid growth of real-time, high-volume communication on services such as Twitter has led to interesting new challenges and opportunities in tracking consistent topics across real-time streams of messages. This paper tackles the problem of tracking topics in a continuous stream of short natural language texts. Our specific application is the popular microblogging service Twitter, where by users can send short, 140-character messages, called "tweets", that are distributed to "followers". Twitter has over 200 million users who collectively create over one billion tweets each week. Whenever major news events happen, the service routinely achieves over 4000 tweets per second.

Mining and monitoring social event in social media has attracting roomy exploration interests, such as social event mining, social event discovery and tracking and event progress examples, which generally implicate a single modality such as text knowledge, social media information incorporate unstructured metadata in many modalities. Generality, nearly all consist duty focuses on either textual features or pictures in separation. Limitation have been constant to analyzing these multimodality in route to model multimedia event accessories.

II LITERATURE SURVEY

Tweets are developed for the use of short text message and it is useful for both users and data analyst. Twitter which gets over 400 million tweets in online per day has emerged as a useful source of news, microblogs, evaluations and extra. Subject matter evolution it must robotically locate subject topic modifications and the moments that they appear. "farzindaratefeh and waelkhreich", a survey of techniques for event detection in twitter, Number 1, 2015. we have propose context semantics when conducting text segmentation. And also hot event evaluation model to identify the user interest. Yan Wu proposed system in 2017 a user-interest model based event evolution model, named the HEE (Hot Event Evolution) model. This model considers the changes in the interests of the user during the evolution of hot events but The most obvious drawback of existing methods for text segmentation is that they only consider surface features and ignore the requirement of semantic coherence within a segmentation. This might lead to incorrect segmentations

1. Ambiguous Segmentation
2. Noisy Short Text
3. Ambiguous Type
4. Ambiguous Instance
5. Enormous Volume

Nikolaos D. Doulamis Proposed system in 2016 used a new event detection algorithm suitable for tweets. To address the dynamic nature of tweets messages can not discovering Hot event.Yali et al. Improved the BTM model in 2015, and put forward a Dirichlet process based BTM model to excute the short text topic. This improvement determines for the number of topics automatically, but the quality of the mining topic has not been improved. Yan et.al. proposed a new kind of bilateral word topic model in 2014, known as the BTM model. This directly analyses the entire document for word co-occurrence patterns and effective way to solve the problem of sparse feature of short text. BTM model cannot analyse the distribution of the topic against individual users and there can be scaling problems when analysing the bilateral words in the large documents. Rosen et al. proposed an author topic model known as ATM in 2012, which can get the author's topic distribution. This model is only suitable for long texts not applicable to short text modelling. Zhao et al proposed a Twitter-LDA topic model in 2011. This also considered user interest information but the model is based on external resources to model the text and generates results with low accuracy.

III NEED

One of the most important tasks of text understanding is to discover hidden semantics from texts. Many efforts have been devoted to this field. For instance, named entity recognition (NER) locates named entities in a text and classifies them into predefined categories such as persons, organizations, locations, etc. Topic models attempt to recognize "latent topics", which are represented as probabilistic distributions on words, from a text. Entity linking focuses on retrieving "explicit topics" expressed as probabilistic distributions on an entire knowledgebase. However, categories, "latent topics", as well as "explicit topics" still have a semantic gap with humans' mental world. As stated in Psychologist Gregory Murphy's highly acclaimed book. "Concepts are the glue that holds our mental world together". Therefore, define short text understanding as to detect concepts mentioned in a short text. Understanding short texts is crucial to many applications, but challenges abound. First, short texts do not always observe the syntax of a written language. As a result, traditional natural language processing tools, ranging from part-of-speech tagging to dependency parsing, cannot be easily applied. Second, short texts usually do not contain sufficient statistical signals to support many state of the art approaches for text mining such as topic modeling. Third, short texts are more ambiguous and noisy, and are generated in an enormous volume, which further increases the difficulty to handle them. We argue that semantic knowledge is required in order to better understand short texts. In this work, we build a prototype system for short text understanding which exploits semantic knowledge provided by a well-known knowledgebase and automatically harvested from a web corpus. Our knowledge-intensive approaches disrupt traditional methods for tasks such as text segmentation, part-of-speech tagging, and concept labeling, in the sense that we focus on semantics in all these tasks.. The results show that semantic knowledge is indispensable for short text understanding, and our knowledge-intensive approaches are both effective and efficient in discovering semantics of short texts.

IV OBJECTIVES

a. To observe the prevalence of ambiguity and limitations of traditional approaches to handling in short text.

b. To achieve better accuracy of short text understanding using semantic analyzer.

c. To improve the efficiency of hot event evolution.

d. To facilitate online instant short text understanding.

V ARCHITECTURE

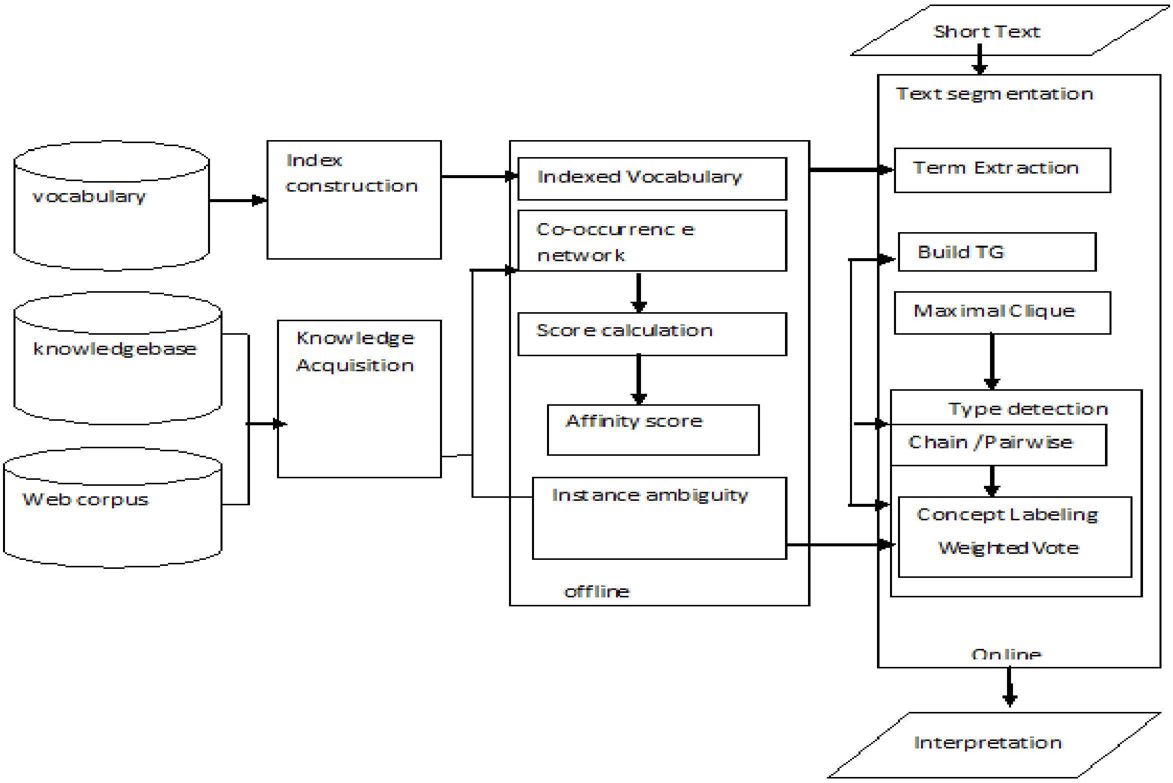

Figure 2. Architecture of proposed system

Although the three steps for short text understanding sound quite simple, challenges still abound and new approaches must be introduced to handle them. In the following, we use several examples to illustrate such a need.

Challenge 1 (Ambiguous Segmentation).

"april in paris lyrics" vs. "vacation april in paris" Both a term and its sub-terms can be contained in the vocabulary,leading to multiple possible segmentations for a short text. However, a valid segmentation should maintain semantic coherence. For example, two segmentations can be derived from "april in paris lyrics", namely fapril in parislyricsg and faprilparislyricsg. However, the former is a better segmentation according to the knowledge that "lyrics" is more semantically related with songs ("april in paris") than months ("april") or cities ("paris").Traditional Longest Cover method, which is widely-adopted for text segmentation , seeks for longest terms contained in a vocabulary. It ignores the requirement of semantic coherence, and thus will lead to incorrect segmentations sometimes. In the case of "vacation april in paris", the Longest Cover method segments it as fvacationapril in parisgwhich is obviously an incoherent segmentation. distance as our similarity function to facilitate approximate term extraction.

VI METHODOLOGY

Indexing of vocabulary and knowledge acquisition.

Approximate term extraction aims to locate substrings in a text which are similar to terms contained in a predefined vocabulary. To quantify the similarity between two strings, many similarity functions have been proposed including token-based similarity functions (e.g., jaccard coefficient) and character-based similarity functions (e.g.,edit distance). Due to the prevalence of misspellings in short texts, we use edit distance as our similarity function to facilitate approximate term extraction.

Text Segmentation.

We can recognize all possible terms from a short text using the tried-based framework described. But the real question is how to obtain a coherent segmentation from the set of terms. We use two examples to illustrate our approachof text segmentation. Obviously, fapril in parislyricsg is a better segmentation of "april in paris lyrics" than faprilparislyricsg, since "lyrics" is more semantically related to songs than two months or cities. Similarly, fvacationaprilparisg is a better segmentation of "vacation april in paris", due to higher coherence among "vacation","april", and "paris" than that between "vacation" and "april in paris".

Type Detection.

Recall that we can obtain the collection of typed terms for a term directly from the vocabulary. For example, term "watch" appears in instance-list, concept-list, as well as verb-list of our vocabulary, thus the possible typed-terms of "watch" are watch[c]; watch[e]; watch[v]g. Analogously, the collections of possible typed-terms for "free" and "movie" are free[ad j]; free[v]g and movie[c]; movie[e]g respectively, as illustrated. For each term derived from a short text, type detection determines the best typed-term from the set of possible typed-terms. In the case of "watch free movie", the best typed-terms for "watch", "free", and "movie" are watch[v], free[ad j], and movie[c] respectively.

Concept Labelling

The most important task in concept labeling is instance disambiguation, which is the process of eliminating in appropriate semantics behind an ambiguous instance. We accomplish this task by reranking concept clusters of the target instance based on context information in a short text(i.e., remaining terms), so that the most appropriate concept clusters are ranked higher and the incorrect ones lower. Our intuition is that a concept cluster is appropriate for an instance only if it is a common semantics of that instance and it achieves support from surrounding context at the same time. Take "hotel california eagles" as an example. Although both animal and music band are popular semantics of "eagles", only music band is semantically coherent (i.e.frequently co-occurs) with the concept song and thus can be kept as the final semantics of "eagles". Text segmentation devide text into a sequence of terms. Statistical approaches that is N-gram Model calculate the frequencies of words co-occuring neighbors in traning corpus. Corresponding neighboring words can be treated as a term threshold are use when frequency exceeds which is predefined.

Vocabulary based approaches extract terms with the help of checking frequency in a predefined vocabulary The most obvious drawback of existing methods for text segmentation is that they only consider surface features and ignore the requirement of semantic coherence within a segmentation. This might lead to incorrect segmentations. I have propose context semantics when conducting text segmentation. And also hot event evalution model to identify the user interest.

POS tagging:-

POS tagging determines lexical types of words in a text. Two types of POS taggers one is Rule-based POS targgers another is statistical POS traggers. Rule based POS traggers attempt to assign POS tags to unknown or ambiguous words that is based on handcrafted or automatically learned linguistic rules. Statistical POS traggers that avoid cost of constructing tagging rules by creating a statistical model.this process done automatically from a corpora and labeling untagged texts based on those learned statistical information.

Semantic Labeling:-

Semantic Labeling are use to disovers hidden semantics from a natural language text. considering representation of semantics. Semantic labeling categorized as namely named entity recognition(NER),topic modeling and entity linking . NER detect and classifies named entities in a short text. Classifies them into predefined categories just like persons, organizations, locations, times, quantities and percentages etc. using linguistic grammar-based techniques as well as statistical models CRF and HMM. "latent topics " are attempt to recognize by topic model which are represented as probabilistic distributions of words. Entity linking provide services to Knowlegebases as a retrieving "explicit topics" which is expressed as probabilistic distributions. High accuracy can be achieve by Semantic Labeling.

Following Algorithms are use

Algorithm-1:- Maximal Clique by Monte Carlo(MaxCMC)

```
Input:
G=(V,E);W(E)={W(e)|eЄE}
Output:
G'=(V',E');S(G')
V'=Ø; E'=Ø
while E≠Ø do
    randomly select e=(u,v) from E with probability proportional to it's weight
    V'=V'U{u,v}; E' U{e}
    V=V-{u,v}; E=E-{e}
    for each t ЄV do
        if e'=(u,t)Є≠E or e'=(v,t)Є≠ E then
            V=V-{t}
            remove edges linked to t from E: E=E-{e'=(t,*)}
        end if
    end for
end while
calculate average edge weight: S(G')=e Є e'∑w(e)/| E'|
```

Algorithm-2: Chunking by Maximal Clique(CMaxC)

```
Input:
G=(V,E); W(E)={w(e)eЄ E}
number of times to run Algorithm 1: K
Output:
G'best=(V'best ,E'best )
Smax=0
for i=1;i<=k; i++ do
    run Algorithm I with  G'i=(V'i,E'i);S(G'i) as output
    if S(G')> Smax then
        G'best= G'i ;  Smax= S(G'i)
    end if
end for
```

VII CONCLUSION

We propose to better understanding and Hot event evolution of short text message in social media. Divide the short text in three categories for better understanding. It increase the effectiveness of short text understanding . chain model and pairwise model are use for type detection. POS taggers is standford Tagger are use for tagging short text. We detail content division as a weighted Maximal Clique issue, and propose a randomized estimation calculation to keep up exactness and enhance proficiency in the meantime.

REFERENCES

[1] Lei-lei Shi, Lu Liu,Yan Wu, Liang Jiang, James Hardy"Event Detection and User Interest Discovering Social Media Data Streams" School of Computer Science and Telecommunication Engineering, Jiangsu University, China 2017

[2] Wen Hua "Understand Short Texts by Harvesting and Analyzing Semantic Knowledge" VOL. 29, NO. 3, MARCH 2017.

[3] Gang Liang, Wenbo He, Chun Xu, Liangyin Chen, and Jinquan Zeng" Rumor Identification in Microblogging Systems Based on Users' Behavior"2016.

[4] Nikolaos D. Doulamis "Event Detection in Twitter Microblogging"*Member, IEEE,* Panagiotis Kokkinos,2016.

[5] Y. Li, C. Jia, and J. Yu, "A parameter-free community detection method based on centrality and dispersion of nodes in complex networks," Physica A: Statistical Mechanics and its Applications, vol. 438, pp. 321-334, 2015.

[6] K. Zhou, A. Martin, and Q. Pan, "A similarity-based community detection method with multiple prototype representation," Physica A: Statistical Mechanics and its Applications, vol. 438, pp. 519-531, 2015.

[7] X. Zhou and L. Chen, "Event detection over twitter social media streams," The VLDB Journal—The International Journal on Very Large Data Bases, vol. 23, pp. 381-400, 2014.

[8] P. Yali, Y. Jian, L. Shaopeng, and L. Jing, "A Biterm-based Dirichlet Process Topic Model for Short Texts," 2014.

[9] X. Zhou and L. Chen, "Event detection over twitter social media streams," The VLDB Journal—The International Journal on Very Large Data Bases, vol. 23, pp. 381-400, 2014.

[10] X. Yan, J. Guo, Y. Lan, and X. Cheng, "A biterm topic model for short texts," in International Conference on World Wide Web, 2013, pp. 1445-1456.

[11] Q. Gao, F. Abel, G.J. Houben, et al., A comparative study of users' microblogging behavior on Sina Weibo and Twitter, in: User Modeling, Adaptation, and Personalization, Springer, Berlin, Heidelberg, 2012, pp. 88-101.

[12] Q. Diao, J. Jiang, F. Zhu, and E.-P. Lim, "Finding bursty topics from microblogs," in Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1, 2012, pp. 536-544.

[13] W. X. Zhao, J. Jiang, J. Weng, J. He, E. P. Lim, H. Yan, et al., "Comparing Twitter and Traditional Media Using Topic Models," Lecture Notes in Computer Science, vol. 6611/2011, pp. 338-349, 2011.

[14] O. Jin, N. N. Liu, K. Zhao, Y. Yu, and Q. Yang, "Transferring topical knowledge from auxiliary long texts for short text clustering," in ACM Conference on Information and Knowledge Management, CIKM 2011, Glasgow, United Kingdom, October, 2011, pp. 775-784.

[15] L. Hong and B. D. Davison, "Empirical Study of Topic Modeling in Twitter," Proceedings of the Sigkdd Workshop on Social Media Analytics, pp. 80-88, 2010.

[16] Linyuan L¨u "Link Prediction in Complex Network"china,2010Arthur Asuncion, Max Welling, Padhraic Smyth On "Smoothing and Inference for topic models," in Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence, 2009, pp. 27-34.

[17] Yanqing Hu "Community Detecting By Signaling on Complex Networks" Department of Systems Science, School of Management Center for Complexity Research Beijing Normal University, Beijing 100875, P.R.China 2008

[18] R. Nallapati, A. Feng, F. Peng, and J. Allan, "Event threading within news topics," in Proceedings of the thirteenth ACM international conference on Information and knowledge management. ACM, 2004,pp. 446–453.

[19] Michal Rosen" The Author-Topic Model for Authors and Documents" Dept. of Computer ScienceUC Irvine,2004

[20] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," Journal of machine Learning research, vol. 3, May. 2003, pp. 993-1022