International Journal of Engineering Research-Online
A Peer Reviewed International Journal
Email:editorijoer@gmail.com http://www.ijoer.in

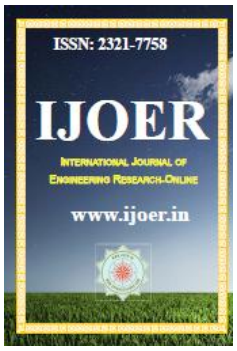
Vol.4., Issue.4., 2016
(July-August)

REVIEW ARTICLE

ISSN: 2321-7758

A REVIEW ON ENTITY LINKING, EXTRACTION, CLASSIFICATION, SENTIMENT ANALYSIS AND LOCATION RECOMMENDATION IN TWEETS

SWATHY.K.SHAJI[1], Dr.SOBHANA N.V[2]
[1, 2] Department of Computer Science and Engineering,
Rajiv Gandhi Institute of technology, Kottayam



ABSTRACT
Twitter has attracted millions of users to share their information creating huge volume of data produced every day. It seems to be a difficult task to handle this huge amounts of data. Here describes a mechanism to extract the valuable information from the tweets using the information extraction techniques. The proposed system describes the entity linking, extraction and classification of data in tweets. As an application of this, here evaluate the performance of sentiment analysis and location recommendation in tweets. The paper mainly focus on the analysis of a number of works related with information extraction and sentiment analysis. The major objective of this review is to provide a more concise and clear idea for the new researches in this area.

Keywords: Twitter stream, Tweet segmentation, Natural language processing, Wikipedia, Named Entity Recognition, Sentiment analysis, Recommendation, Collaborative Filtering

I. INTRODUCTION

Twitter has attracted millions of users to share their information creating huge volume of data produced every day. It is a very difficult and time consuming task to handle this huge amount of data. Social media usually refers to the user generated data such as tweets, Facebook updates, blogs etc. Such data are became immense and many such applications need to perform the entity linking, extraction and classification of data. Suppose consider a string "Obama gave an immigration speech in Hawaii", entity extraction actually determine that the string Obama refers to a person name and Hawaii refers to a location. The entity linking actually do the task by inferring the entity "Obama" with an external Knowledge base for example: en.Wikipedia.org/wiki/Barack_Obama[1].

Named entity recognition is the task of identifying the named entities. The entity can be a person name, an organization name, location name,percentage value etc. Named entity recognition and text classification are well known problems in natural language processing. The paper describes an end to end industrial system that extracts, links and classify the entities in tweets. Here the entities are being linked by using an external knowledge base. The paper uses a global real time knowledge base called the Wikipedia. Wikipedia is global and may contain several concepts and instances[1]. The real time nature of Wikipedia makes it well suitable for handling the social data. The paper describes four tasks: entity extraction, linking, classification and tagging of social media data. The paper discusses all the tasks related with information extraction, sentimental analysis and location recommendation.

Another way of identifying the entities is by the way of using social contexts and social

information. For example consider a string having a name "Mel Gibson"[1], when using the global real time knowledge base Wikipedia it is difficult to identify the name. By using the local context information the tweets during an hour can be grouped and to determine whether the name is used in more than one tweets. If it is specified in more tweets then it should be considered as an entity and can be easily identified. The system architecture has been depicted by the figure below [1]:

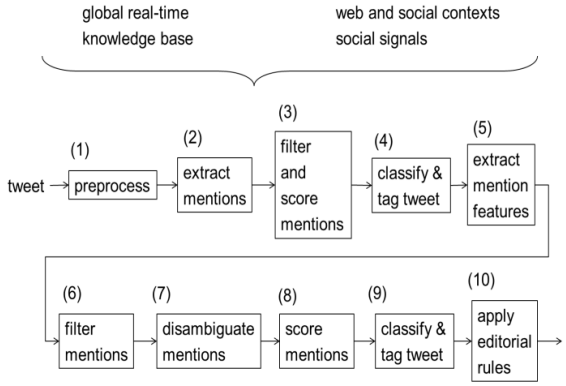


Figure.1. System Architecture

Sentiment analysis is another emerging area in social media. Sentiment analysis evaluate the tweets to find whether the tweets are positive, negative or neutral. It is a way of determining the sentiment tendency towards a topic without reading the whole tweets. As an application of this information extraction here describes a location recommendation system that recommend the places of interest to the target users based on the user's interest. Tweet users have the provision to rate the tourist places and based on their rate of interest the new places can be recommended. The recommendation system actually uses the collaborative filtering algorithm for recommending the place of interest to the target users.

II. RELATED WORKS

Agichtein [2] et.al proposes the mining reference tables for automatic text segmentation. The paper which exploit the reference relations that relates with the clean tuples. It exploit the widely available reference tables in most data warehouses to built a robust segmentation system. Here proposes a CRAM [2] system that is basically a two-phased approach. In the first pre-processing phase here built a attribute recognition model over each column in the reference table. In the second segmentation phase, a string s is being segmented to its constituent values s1.....sn and then assign each si to a distinct column.

In the segmentation model, here evaluate a attribute recognition model that correctly segment the string into its corresponding constituent components such as person name, organization name, street address, state etc. For example given a string s, the segmentation problem should partition s into s1.....sn and to map these into some distinct attributes. Here describes an HMM [2] which requires a topology that consists of a set of states, transitions and also the emission probabilities and the transition probabilities between the states. The set of states in an ARM [2] model is categorized into beginning, middle and trailing states. Each of these category consists of state s for each element e in the corresponding categorized dictionary.

In order to complete the instantiation of the attribute recognition mode, it is needed to define the emission probabilities of each states and also the transition probabilities between the states. The paper describes a segmentation algorithm for effectively segmenting the strings. The segmentation procedure may consists of two components. Firstly , it is needed to determine the sequence in which attribute values are concatenated and should effectively determine the best segmentation method for segmenting the string into its attribute values. Thus the paper exploits the reference tables for segmenting the strings into structured records.

Aitken [3] proposed a paper for learning information extraction rules by applying the inductive logic programming technique. Here uses the FOIL-LP [3] learner and the problem result in an appropriate representation of the text.

The inductive logic programming is appropriate for learning task since it provides a natural representation of the relations. The paper describes an ILP [3] algorithm, and the input to it consists of two parts, First a list of positive and negative instance of relations and second the background

theory from which rules are constructed. In the NLP tasks, POS tags can be used as a filtering mechanism. The system uses several natural language processing tasks such as POS tagging, morphological analysis, POS tag convergence, pos filtering, Frequency analysis, named entity recognition etc.

The paper mainly focuses on the ontology theory and the isa(Class,Class) relation should be considered as a background theory. Named entity recognition task results in a has word relation which specifies that a concept or entity occur in a sentence. In the analysis phase, here evaluate the precision, recall and F scores. Precision is actually the ratio of the correctly derived relations to that of the total number of derived relations. Recall is the ratio of the number of correctly derived relations to that of the total number of correct relations. F score has been calculated by giving equal weight for both precision and recall.

Douglas [4] et al proposes a finite state processor for information extraction from the real world text. Normal text processing with the basic task of parsing text is tend to be slow and error prone but the FASTUS [4] which is a non deterministic finite state language model that effectively provides a phrasal decomposition of string into noun phrases, verb phrases and the particles. FASTUS is therefore a system that provides better accuracy of extracting the prespecified information from the text. The system has been tested in MUC_4 [4] evaluation of text processing system and has bring the following advantages:

(1) High performance in terms of recall and precision values.

(2) Has short domain specific development time.

(3) Very fast processing time.

The paper describes that the natural language system can be broadly classified as information extraction system and text understanding system. The information extraction system has to deal with the fact that only a portion of text is relevant. In this task the information is mapped into predefined relatively simple categories. In the MUC_4 evaluation [4] the principal measures of evaluation are recall and precision. Recall which actually describes the number of right answers the system got divided by the total number of right answers. Precision which brings the measure of the number of right answers the system got to the total number of answers the system have.

Here evaluate a FASTUS [4] system architecture. The text is first preprocessed o obtain the remainder of the processing. The preprocessing phase which brings a spelling correction as well. The major operations of FASTUS [4] composed of four steps: (1) triggering (2) recognizing the phrases (3) recognizing the patterns (4) merging incidents. In the triggering phase, the triggering words should be searched. Generally these triggering words are the least frequent words required by the patterns. Recognizing the phrases deals with identifying the noun phrases by parsing the text. The noun group were recognized by the 37-state non- deterministic finite state automation. The verb groups are being determined by the 18 state non deterministic finite state machine. Next the task of recognizing patterns, the patterns are encoded as finite state machines where the state transition are effected. Here they implement 95 patterns for MUC-4 [4] application. Merging incidents phase has to merge the incidents with others in a sentence. The advantages of FASTUS [3] system are:

(1) Relatively simple

(2) Basic system is relatively small

(3) Very effective

(4) Fastest run time

(5) Fast development time

Vinayak [5] et al proposed a mechanism of segmentation of text into structured records. The paper present a method to automatically segment the unformatted text records into some structured elements. The experiments shows that it brings 90% accuracy in Asian addresses and 99% accuracy in US addresses. Here the task is to extract the semantic entities from documents.

Paper proposes a tool called DATAMOLD [4] for automatically segmenting such data. DATAMOLD [5] is basically a technique with the features of Hidden Markov Modeling (HMM) [5]. HMM [5] is a probabilistic finite state automation which comprise of a set of states, dictionaries of discrete output symbols and a set of transitions from one state to

another state. Here there is two states namely the start state and end state. On beginning from start state the HMM which produce a number of output sequence can be attained from multiple path each with its own probabilities. The paper presents a nested model for learning in HMM. The outer HMM [5] focuses on the sequence relationship between elements and the inner HMM [5] which learns the finer structure within each elements. Paper shows a mechanism to integrate an external database into HMM model[5]. The system which identifies all the elements in text and should be merged into some specific categories.

The input to a DATAMOLD [5] is a fixed set of elements having 'house,'street', 'city' etc. The output sequence can be generated from multiple paths. The multiple paths should have different probability values. Here Viterbi approximation can be applied and by this we can say that the path having highest probability should be the output sequence. The paper evaluated the system performance by using three datasets namely US addresses, Student addresses and company addresses. Text segmentation task is very difficult in the sense that the data is highly irregular.

Califf [6] et al proposes a system that focuses on the relational learning of pattern matching rules for information extraction. Information extraction system processes the documents to reference a specific set of relevant items [6]. Here proposes a system named RAPIER [6] (Robust automated products of information extraction rules), it learns the rules for the information extraction task. The algorithm that learn from inductive logic programming system (ILP)[6] [5]. Here the major focus is to extract information from Usenet newsgroups. RAPIER [6] rule representation use patterns that have syntactic and semantic information. Extraction rules are being indexed by a template name and slot name, it must have three different parts: (1) pre-filler patterns (2) pattern that match with the actual slot filler. (3) post filler pattern. RAPIER [6] attempt to compress and generalize the rules for each slot. New rules are being created by selecting two already existing rules and generalizing them. If the best rule which produce no negative examples then it should be under consideration and that rule is being added to the rule base.

Xiaolong Wang [7] et al proposed an approach for obtaining the common sentiment tendency towards a topic without reading the whole tweets. The paper mainly focus on hash tag level sentiment classification [7] in tweets. The paper describes three types of information that is relevant to the task they are:

(1) Sentiment polarity of tweets having hash tag [7].

(2) Hash tag co-occurrence relationship.

(3) Literal meaning of hash tags.

Here proposes a model graph [7] for classification and also evaluate three approximate collective classification algorithms. Hash tags described here are community driven convention for adding additional context to tweets. Hash tags [7] are created organically by grouping the messages and to find the topics that highlight in the text. The hash tag is represented by prefixing a word with symbol # (eg. '#hashtag'). Thus the extensive use of hash tags [7] in tweets are very expressive and attracts the twitter users. The basic idea behind the approach is to aggregate the sentiment polarity with the classification results for each of the tweets that has a hash tags [7]. The paper mainly relies its focus on sentiment polarity classification based on hash tags and not on the way of sentiment analysis in tweets.

Tim Finn [8] et al presents the idea that annotation techniques will provide the first step towards the full study of named entities in social networks. Here describes the fact that in the twitter community the symbol '@' character at the beginning indicate that it is a message directed by the user and the @ character at the middle indicates that it is a general reference to the user. The paper describes the experience of doing task using Amazon mechanical Turk [8] and Crowd flower [8]. Separate task has been developed on Crowd flower [8] and MTurk [8] using collection of twitter status in order to determine which will perform well. MTurk has an advantage of using standard HTML and JavaScript. But the MTurk has inferior data verification Crowd Flower works across multiple services and does its verification with gold standard data [8]. The twitter

messages are analyzed to assign toggle button to tag words with person (PER), Organization(ORG), Location(LOC) etc. Person (PER) indicates the name of person, nick name or alias. Organization (ORG) deals with Institution, government agencies, Corporation etc and the Location (LOC) usually denotes the geographically defined places, cities, States etc.

The four basic principles used by the annotator while tagging were:

(1) Tag words according to their meaningful

(2) Only tag words that refers to entities.

(3) Only tag names of type PER, ORG and LOC.

(4) Use ??? symbol to indicate the uncertainty.

Here they evaluated the gold standard dataset of about 400 tweets.

John Lafferty [9] et al propose the conditional random field , a framework for building probabilistic models for segmenting the sequence data. The Conditional random field has many advantages over hidden Markov model and Stochastic grammars. Here present an iterative parameter estimation algorithm for conditional random field and then compare the performance with both HMM and MEMM (Maximum Entropy Markov model). HMM and stochastic grammars has been used for a wide variety of problems in text including part of speech tagging, topic segmentation, information extraction etc.

In Maximum entropy Markov models, each of the source state has a exponential model and which takes as input the observational features and outputs the distribution over next states. The exponential models are being trained by an iterative scaling mechanism in the maximum entropy framework. Here describes a weakness dealt with the discriminstic Markov model called the label bias problem.

Alexis Mitchell [10] et al presents four challenges that were met in the field of information extraction and they were:

(1) Recognition of entities

(2) recognition of relations(generally has 5 types of relations and may include role, part, At, Neck, Social)

(3) Event extraction

(4) Extraction which is measured not only on text but also on speech recognition

The paper describes the named entities, relations and events in terms of their attributes and constituents. Entity attributes are usually of the type person, Organization, Location etc. Entity relations are being represented in terms of attributes and its arguments. Similarly events are being represented in terms of attributes and its participants. Some examples regarding the events were destroy, create, move etc. The entities thus described can be further being tagged to some class – Specific, Attributive, negatively quantified, generic or underspecified.

Li Weigang [11] et al proposed a unified approach for domain specific tweet sentiment analysis. The paper proposes a unified tool named UnB TSA [10] which consists of 4 steps including tweet collection, refinement, sentiment lexicon creation and sentiment analysis. Tweet collection actually dealt with the way of collecting tweets that are related with a specific topic. Refinement should include the way of selecting relevant tweets. Third step sentiment lexicon creation is essential to capture the specificity. Final task is performing the sentiment analysis.

The paper proposed a TSA tool [11] had been tested on product iphone6 and has obtained good results in all four phases.

Rabia Batool [12] et al proposes a precise tweet classification and sentimental analysis system. Twitter has enormous data and to extract valuable information from the huge messages is a difficult task. Here the paper describes a mechanism to classify the tweet and sentiment based on the data it contain. The information contained in the tweets are being extracted using keyword based information extraction mechanism. The keywords, entities thus extracted can be used for the tweet classification and sentiment analysis tasks.

The paper applies a knowledge enhancer and synonym binders module on these extracted information and thus the information gain can be increased by 0.1 to 55%. The research paper discusses a system to process short text and to filter them effectively. The system analyzes the tweets for extracting the features and thus enhance its

usability in sentiment analysis. Here for increasing the information gain the system applies filtering on keywords, entities etc.

The proposed system which uses twitter data and it performs the task of parsing, domain specific classification and sentiment analysis. There are a number of components that enable the processing and analysis of tweets:

1) PreProcessor: Tweets sholud be preprocessed to extract the information. DOM parser is being used for parsing the text. Parser split data into username, tweetdata, status, Tweetid and image.

2) Knowledge generator: The generator extract the valuable information from tweets and classify tweets into different categories. It actually accept unstructured text and process it using natural language processing and machine learning techniques. The system which extracts the keywords and their associated sentiments.

3) Knowledge Enhancer: The module add additional knowledge and the system which uses part of speech tagging and entity extraction on tweets.

4) Synonym Binder: It is an additional step to increase the information gain from tweets. Wordnet dictionary has been used to bind synonym with entities.

5) Filter Engine: Filtering has being applied on the information extracted for classifying the tweets into different categories. Filter engine filters the tweet and stores them in a repository.

III CONCLUSION

In this paper, we have made an analysis of the different techniques related with the information extraction. The information extraction from Wikipedia has accursed importance now a days. As Wikipedia contain huge amounts of data, extracting the valuable information from this is considered to be a tedious task. The paper mainly discusses the process of entity linking, extraction and classification of data in tweets. The work analyze some of the already developed systems in information extraction and sentiment analysis. This review on information extraction will definitely provide more clear and concise idea for the new researchers.

REFERENCES

[1]. Abhishek Gattani , Digvijay S. Lamba , Nikesh Garera , Mitul Tiwari , Xiaoyong Chai et al. Entity Extraction, Linking, Classification, and Tagging for Social Media: A Wikipedia-Based Approach. The 39th International Conference on Very Large Data Bases, August 26th - 30th 2013, Riva del Garda, Trento, Italy.

[2]. E. Agichtein and V. Ganti. Mining reference tables for automatic text segmentation. In SIGKDD, 2004.

[3]. J. S. Aitken. Learning information extraction rules: An inductive logic programming approach. In ECAI, 2002.

[4]. D. E. Appelt, J. R. Hobbs, J. Bear, D. Israel, and M. Tyson. FASTUS: A finite-state processor for information extraction from real-world text. In IJCAI, 1993.

[5]. V. Borkar, K. Deshmukh, and S. Sarawagi. Automatic segmentation of text into structured records. In SIGMOD Record, 2001.

[6]. M. E. Califf and R. J. Mooney. Relational learning of pattern-match rules for information extraction. In AAAI, 1999.

[7]. Xiaolong Wang, Furu Wei, Xiaohua Liu, Ming Zhou, and Ming Zhang. Topic sentiment analysis in Twitter: A graph-based hashtag sentiment classification approach. 20th ACM Conference on Information and Knowledge Management, CIKM 2011, Glasgow, United Kingdom, October 24-28, 2011

[8]. T. Finin, W. Murnane, A. Karandikar, N. Keller, J. Martineau, and M. Dredze. Annotating named entities in Twitter data with crowdsourcing. In HLT-NAACL, 2010.

[9]. J. Lafferty, A. McCallum, and F. C. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In ICML, 2001.

[10]. G. Doddington, A. Mitchell, M. Przybocki,nL. Ramshaw, S. Strassel, and R. Weischedel. The automatic content

extraction (ACE) program–tasks, data, and evaluation. In LREC, 2004.

[11]. Patricia L V Ribeiro and Li Weigang, Tiancheng Li. A Unified Approach for Domain-Specific Tweet Sentiment Analysis. In 18th International Conference on Information Fusion Washington, DC - July 6-9, 2015.

[12]. Rabia Batool, Asad Masood Khattak, Jahanzeb Maqbool and Sungyoung Lee. Precise Tweet Classification and Sentiment Analysis. 2013 IEEE.