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

Vol.6., Issue.1, 2018
Jan-Feb

RESEARCH ARTICLE

ISSN: 2321-7758

ENTITY LINKING, EXTRACTION, SENTIMENT ANALYSIS AND LOCATION RECOMMENDATION IN TWEETS

SWATHY.K.SHAJI, Dr.SOBHANA N.V
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 amount 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 and also hybrid approach using CRF. As an application of this, here evaluate the performance of sentiment analysis and location recommendation in tweets. The major objective of this is to provide a more concise and clear idea for the new researches in this area.

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

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 Knowledgebase 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. The system architecture has been depicted by Fig.1.[1] as below:

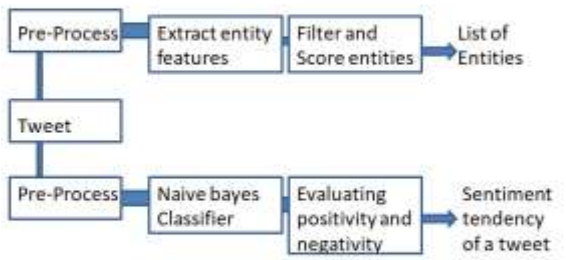

Fig.1. System Architecture

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 recommends 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.

1. 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. 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. 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. 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. 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]. 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]. RAPIER [7] rule representation use patterns that have syntactic and semantic information. Xiaolong Wang [8] et al proposed an approach for obtaining the common sentiment tendency towards a topic without reading the whole tweets. Tim Finn [9] et al presents the idea that annotation techniques will provide the first step towards the full study of named entities in social networks. John Lafferty [10] 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. Alexis Mitchell [11] 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. Li Weigang [12] et al proposed a unified approach for domain specific tweet sentiment analysis. Rabia Batool [13] 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.

2. NAMED ENTITY RECOGNITION

Named entity recognition is the task of recognizing the proper nouns or entities in text and relating them to a predefined set of categories.

Most of the NER systems are based on analyzing the patterns of POS tags. The categories to which the entities belong may be location name, organization, date, expression, percentage, person name etc. Named entity recognition [4] has a number of applications in the field of natural language processing. It has been used in information extraction, parsing, machine translation, question answering etc. NER systems have been mostly used in the field of BioInformatics and molecular biology for extracting the entities. Most of the NER systems are word based, here we employ the segmentation method for the identification of named entities. Named entity recognition task has been performed for evaluating the performance of two algorithms: Random walk model and POS Tagging.

3. Sentiment Analysis

Sentiment analysis or opinion mining is the study of analyzing one's opinion, sentiment, attitudes, emotions towards events, topics, products etc. Research in sentiment analysis has a potential impact on economic, social and political scenarios. In the real world, sentiment analysis has more impact on social media. For example, if one wants to buy a product, he will definitely look at the user reviews and discussions in web. Sentiment analysis application has been developed in many fields such as consumer products and services, financial services, healthcare and poltical election. Nowadays sentiment analysis technique can be used for finding the election result based on people's opinions.

3.1 Naive Bayes Algorithm

Naive Bayes is a popular algorithm for classifying the text. Bayes classifier can be used for counting the number of appearances of words, documents and categories. It can also be combined for evaluating the probability of each classes. Naive Bayes classifier is one of the most useful machine learning technique. Let us consider an example of finding the probability that a tweet can be classified as positive or negative. Let P(A) which represents the probability of class and P(B) represents the probability of tweets. Then P(B/A) denotes the probability of tweet B given Class A. Similarly we can denote P(A/B) as the probability of class A given tweet B.

The Naive Bayes theorem evaluates the probability as below:

$$P(A/B)= \frac{(P(B/A)*P(A))}{(P(B))} \quad (2)$$

In this paper we are evaluating the relative probabilities of a sentence being positive, negative or neutral. Let Positivity denotes the probability value of a sentence being positive and Negativity denotes the the probability value of a sentence being negative. Therefore positivity can be evaluated by using the equation as below:

$$\text{Positivity}= \frac{(Positivity)}{(Positivity+Negativity)} \quad (3)$$

and

$$\text{Negativity}= \frac{(Negativity)}{(Positivity+Negativity)} \quad (4)$$

Hence the measure of positivity and negativity can be used for evaluating whether the tweets are positive, negative or neutral.

4. APPLICATIONS

4.1 Recommendation System

Recommender system is important field of information sharing and e-commerce. It is a powerful method to filter through large information databases. Many of the e-commerce websites has been using this recommendation system for recommending the product of interest to the targeted users. For example, in the case of Amazon.com they have been using collaborative filtering for recommending the items of interest to the intended users. Collaborative filtering system consists of users and their preferences called ratings for certain items. The preference expressed by a user for a particular item is known as rating. The triplet that is formed by a collaborative filter is (user,item,rating).

User-User Collaborative Filtering

It is also called by the name K-NN Collaborative Filtering. It is a collaborative filtering technique in which we find out whether the other users past rating behavior are similar to that of the current user.

User based CF

In this work we used the user based collaborative filter for computing the similarity between users and to recommend the locations to the target users. The collaborative filtering uses a database of preferences made by users for locations. User based CF is a personalized recommendation framework that recommend the locations to a user by examining the preferences made by the user and the other users. In other words it can be defined as a way of predicting the rating behavior of the current user based on the preferences made by the other users. Basically this is an aggregation of K nearest neighbor. Suppose there are K users who were nearest to a user, then it can be termed as K nearest neighbor. The similarity between the users can be found out by using two methods: correlation and cosine method. This paper discusses about the pearson correlation for finding the similarity between the users. Pearson correlation uses the following formula for finding the linear correlation between the users [13]:

$$Sim\,(m,n) = \frac{\sum_{a \in I_{mn}} (R_{m,a} - A_m)(R_{n,a} - A_n)}{\sqrt{\sum_{a \in I_{mn}} (R_{m,a} - A_m)^2 \sum_{a \in I_{mn}} (R_{n,a} - A_n)^2}} \qquad (5)$$

Where R_{mn} is the rating of location a by user m. Am is the average rating of user m for co-rated locations. I_{mn} is the location set rated by both user m and n.

For the selection of neighbors a threshold selection method is being used. According to this method if the similarity between users exceeds a certain threshold then they should be considered as neighbors.

The rating (p_{xk}) of target user k to a location x is as below [13]:

$$P_{xk} = \bar{A_x} + \frac{\sum_{m=1}^{C} (R_{mk} - A_m)^{*} Sim(x,m)}{\sum_{m=1}^{C} Sim(x,m)} \qquad (6)$$

Where A_x is the average rating of user x for the locations. R_{mk} is the rating made by the user m for the location k. Am is the average rating made by the user m for the locations. Sim(x,m) is the similarity measure between the user x and its neighboring user m. C represents the total number of neighbors.

5. EXPERIMENTAL EVALUATION

5.1 Experimental Settings

Experiment on tweet classification has been done by using tweet dataset having 500 English tweets. Experiment has been done by using the POS tagging method. Here also used Wikipedia for evaluating whether the entity belongs to the category of person, organization or location. Wikipedia dump is also evaluated for comparing the entities, ie usually contains 3,246,821 articles in whole and there should e about 4,342,73 distinct entities appeared as anchor text in these articles. Tweet segmentation is being used in tweet classification techniques along with POS tagging. The tweet dataset evaluated here contain the fields author name, date published and the tweets. The date published is being used for grouping the tweets under a certain publication period by using the concept of learning from local context.

A NE tagged twitter corpus has been used for NER experiment. This corpus is split into two sets. One forms the training data and the other forms the test data. They consist of 90% and 10% of the total data respectively. CRF is trained with training data and test data is tagged using CRF model. More than 2 lakh words have been used as training set for the CRF based NER system. The size of the test file is 23K words and the data is labelled with 17 labels. We have considered features such as prefix and suffix of length up to three of the current word, POS information, digit features, information about the surrounding words and their tags.

We have used different standard measures such as Precision, Recall and F-measure for evaluation. Recall is the ratio of number of NE words retrieved tothe total number of NE words actually present in the file (gold standard). Precision is the ratio of number of correctly retrieved NE words to the total number of NE words retrieved by the system. These two measures of performance combine to form one measure of performance, the F-measure, which is computed by the weighted harmonic mean of precision and recall.

5.2 Performance Measure

The proposed system has three different frameworks one is based on entity linking, extraction and classification using Wikipedia, the other one deals with the sentimental analysis on tweets based on positive and negative dataset and by linking with Wikipedia. This is the first paper that extract the movie reviews from different review websites for predicting whether the sentiment is positive, negative or neutral. The third framework used is the location recommendation framework that is entirely based on collaborative filtering. The

International Journal of Engineering Research-Online
A Peer Reviewed International Journal
Articles available online http://www.ijoer.in; editorijoer@gmail.com
Vol.6., Issue.1, 2018
Jan-Feb

collaborative filter will recommend the location of interest to the targeted users.

Performance analysis can be done by using different measures of analysis methods like precision, recall and F-measure. In information retrieval contexts, precision and recall are defined in terms of a set of retrieved documents and a set of relevant documents.

Recall in information retrieval is the fraction of the documents that are relevant and are successfully retrieved. F measure can be considered as a balanced score by using precision and recall values. Table.1. shows the precision, recall and F measure value for the identification of person name, location name and organization name.

TABLE.1. PRECISION VALUE FOR THREE NAMED ENTITIES

Name of Entity	**P**	**R**	**F**
Person name	0.71	0.826	0.771
Location name	0.82	0.85	0.824
Organization name	0.72	0.672	0.571

The values of precision, recall and F measure for the three named entities can be plotted using a graph that correctly differentiates the variation of each of the values for the three named entities. Figure.2. which shows the variation of precision, recall and F measure for three named entities person, location and organization.

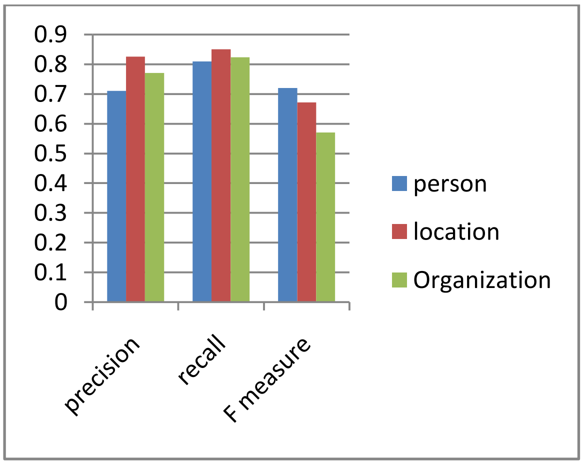

Fig.2. – Precision, Recall and F measure for the three named entities person, location and organization

7.4 Comparison with Existing Methods

We compare our system with the popular Stanford Named Entity Recognizer and the popular industrial system OpenCalais. Here we consider three versions of the Stanford system:

1) StanNER-3: This is a 3-class (Person, Organization, Location) named entity recognizer. The system uses a CRF-based model, trained on a mixture of CoNLL, MUC and ACE named entity corpora.
2) StanNER-3-cl: This is the caseless version of StanNER-3 system, it ignores capitalization in text.
3) StanNER-4: This is a 4-class (Person, Organization, Location) named entity recognizer for English text. This system uses a CRF-based model which trained on the CoNLL corpora.
4) OpenCalais is an industrial product of Thomson Reuters which provides open APIs to extract entities, facts, events and relations from text .

The table 2,3,4 show that our system outperforms the other two in almost all aspects, especially with respect to extracting organizations. A main reason for low precision in the other systems is that they interpret many interjections (ro,lmao, haha, etc) and abbreviations as organization names. A main reason for low recall is the difficulty in recognizing an organization name without using a large KB. For example, most NER tools without a large KB would incorrectly identify “Emilie Sloan" as a person, not an organization.

TABLE.2. MEASURES FOR PERSON

Method	P	R	F
System using Entity linking	0.71	0.826	0.771
System using CRF	0.7	0.76	0.69
System using Hybrid Approach	0.69	0.78	0.75
StanNER-3	0.69	0.42	0.54
StanNER-3cl	0.70	0.56	0.69
StanNER-4	0.69	0.56	0.59
OpenCalais	0.67	0.43	0.56

International Journal of Engineering Research-Online
A Peer Reviewed International Journal
Articles available online http://www.ijoer.in; editorijoer@gmail.com
Vol.6., Issue.1, 2018
Jan-Feb

TABLE.3. MEASURES FOR LOCATION

Method	P	R	F
System using Entity linking	0.82	0.85	0.824
System using CRF	0.81	0.80	0.793
System using Hybrid Approach	0.82	0.843	0.83
StanNER-3	0.71	0.50	0.69
StanNER-3cl	0.56	0.70	0.65
StanNER-4	0.67	0.43	0.56
OpenCalais	0.68	0.44	0.57

TABLE.4. MEASURES FOR ORGANIZATION

Method	P	R	F
System using Entity linking	0.71	0.672	0.571
System using CRF	0.70	0.663	0.556
System using Hybrid Approach	0.712	0.667	0.564
StanNER-3	0.47	0.10	0.17
StanNER-3cl	0.60	0.12	0.23
StanNER-4	0.21	0.26	0.23
OpenCalais	0.42	0.09	0.15

Our System vs OpenCalais:

Open Calais is quite similar to our system, in that it can perform all four tasks of extraction, linking, classification, and tagging, and that it can handle a large number of categories (in contrast, the current Stanford variants only focus on extracting persons, organizations, and locations).

7.5 Sentiment Analysis

Sentiment analysis accuracy can be evaluated by determining whether the positive, negative or neutral sentiment is correctly classified or not. The sentiment analysis task can be evaluated using its performance by using the positive and negative datasets. Here the sentiment analysis is also evaluated using the Wikipedia reviews and other review websites. The performance of sentiment analysis can also be evaluated by using the precision, recall and F1 measures.

It can be represented by using the Table.5. as below:

TABLE.5. PRECISION, RECALL AND F MEASURE VALUE FOR THREE SENTIMENTS

Type of Sentiment	**P**	**R**	**F**
Positive	0.332	0.80	0.469
Negative	0.302	0.78	0.435
Neutral	0.324	0.72	0.446

Similarly the values of precision, recall and F measure for the three sentiments can be plotted using a graph that correctly differentiates the variation of each of the values for the three sentiments. Fig.3. which shows the variation of precision, recall and F measure for three sentiments positive, negative and neutral.

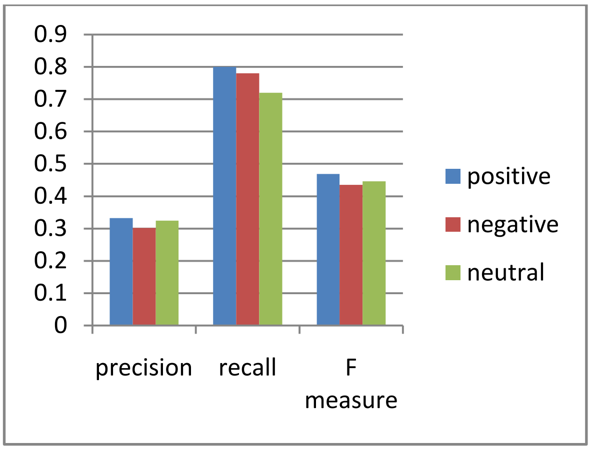

Fig.3. Precision, Recall and F measure for positive, negative and neutral sentiments.

7.6 Comparison with other Classifiers

The most frequently used machine learning classifiers are naïve bayes, Maximum Entropy (MaxEnt), and Support Vector Machines (SVM). Since the training process treats emoticons as noisy label, we have to remove the emoticons from the tweets. If we remove the emoticons it would have a negative impact on both Maximum Entropy and support vector machines, but less impact on Naïve Bayes. This is because of the difference in the mathematical models.

The idea behind maximum entropy model is that it would prefer the most uniform model that satisfies by given constraint. It is a feature extracted model. Support vector machine classifier is another popular classification method. The accuracy obtained for our classifier on comparison with other classifiers is indicated as below in table.6:

International Journal of Engineering Research-Online
A Peer Reviewed International Journal
Articles available online http://www.ijoer.in; editorijoer@gmail.com
Vol.6., Issue.1, 2018
Jan-Feb

TABLE.6. ACCURACY OF DIFFERENT CLASSIFIERS

Type of sentiment	Naïve Bayes	MaxEnt	SVM
Positive	83.3	80.4	80.9
Negative	84.2	83.6	81.7
Neutral	78.9	76.4	78.5

Graphical representation of the accuracy of different classifiers can be represented by the below fig.4.

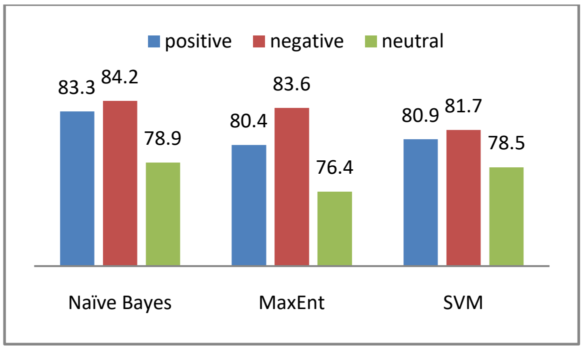

Fig.4. Accuracy of different classifiers

7.7 Location Recommendation

Location Recommendation system can be evaluated for its performance by evaluating the similarity between the users in the neighborhood of the current user. Here used the Pearson correlation for finding the similarity between the users. The similarity between different users based on their rating for the locations can be indicated as below:

TABLE.7. SIMILARITY MEASURE BETWEEN USERS

First user	Second user	Similarity
User#1	User#2	0.5
User#2	User#3	0.309
User#4	User#2	0.333
User#2	User#5	0.167
User#6	User#2	0.85
User#2	User#7	0.333

In the paper we use the statistical accuracy for evaluating the performance of the Recommendation system. The statistical accuracy metrics determine the Mean Absolute Error (MAE) [14] by using the formula:

$$MAE = \frac{\sum_{i=1}^{n} p_{i-q_i}}{n} \qquad (7)$$

Where p_1,……p_n determines the predicted ratings and q_1,…..q_n denotes the actual ratings. The prediction value should be maximum for minimum valued MAE and it can be easily understand by using the graph as below:

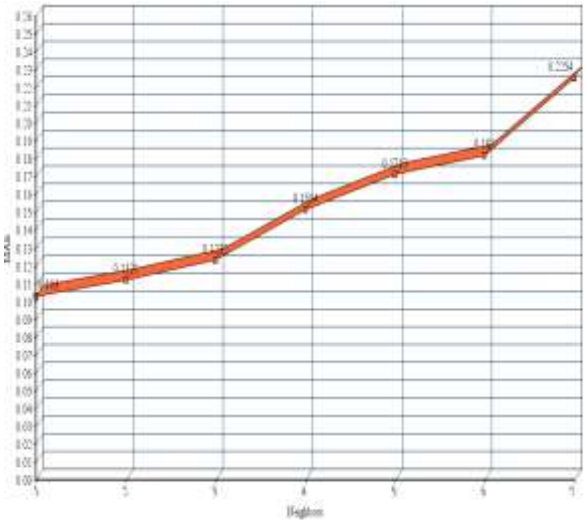

Fig.5. Graph showing the MAE value for different number of nearest neighbors

6. Conclusion

In this paper, we have made an analysis of entity linking, extraction and classification of data by using the Wikipedia content. The information extraction from Wikipedia has acquired importance now 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, sentiment analysis and location recommendation in tweets. We have also proposed hybrid approach using CRF. This proposed method has obtained better accuracy than CRF and entity linking methods. This is the first paper that describes the sentimental analysis based on the data from review sites and by using the positive and negative datasets.

REFERENCES

[1]. Abhishek Gattani , Digvijay S. Lamba , Nikesh Garera 1 , Mitul Tiwari 2 , 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.J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[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. M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

[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.

[13]. Li Ya Li ; Pan Shan Liang ; Xuang Xi " A Kind Of Web Service Recommendation Method Based On Improved Hybrid Collaborative Filtering", *International Conference on Cognitive Informatics & Cognitive Computing(ICCI*CC),2012.*

[14]. Hengsong Tan, Hongwu ye," A Collaborative filtering recommendation Algorithm based on item Classfication ", IEEE Pacific Asia Conference on Circuits, Communication and system, China, 2009, pp 694-697.

[15]. www.google.com

[16]. Wakao, T., Gaizauskas, V. and Wilks, Y. 1996. Evaluation of an algorithm for the recognition and classification of proper names, In Proceedings of COLING-96.

[17]. Singh, A. K. and Surana, H. 2007. Can Corpus Based Measures be Used for Comparative Study of Languages, In Proceedings of Ninth Meeting of the ACL Special Interest Group in Computational Morphology and Phonology, ACL 2007.

[18]. Bikel, D. M. , Schwartz, R. L. and Weischedel R. M. 1999. An Algorithm that Learns What's in a Name. Machine Learning, pp. 211-231.

[19]. Borthwick, 1999. Maximum Entropy Approach to Named Entity Recognition, Ph.D. thesis, New York University.

[20]. Lafferty, J. D., McCallum, A. and Perera, F. C. N. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data, pp.282-289, ICML 2001.

[21]. Li W and McCallum A. 2003. Rapid development of Hindi named entity recognition using conditional random fields and feature induction, ACM Transactions on Asian Language Information Processing (TALIP), pp.290–294.

[22]. Srihari, R., Niu, C. and Li, W. 2000. A Hybrid Approach for Named Entity and Sub-Type Tagging, In Proceedings of the sixth conference on Applied natural language processing.

[23]. Toral, A., Noguera, E. Llopis, F. and Munoz, R. 2005. Improving question answering

using named entity recognition, In Proceedings of the 10th NLDB congress, Lecture notes in Computer Science.

[24]. Spain, A., Babych B. and Hartley. 2003. A. Improving machine translation quality with automatic named entity recognition, Springer-Verlag 2003.

[25]. Wallach, H. M. 2004. Conditional random fields: An introduction, Technical Report MS-CIS-04-21, University of Pennsylvania, Department of Computer and Information Science, University of Pennsylvania.