International Journal of New Technology and Research (IJNTR)
ISSN:2454-4116, Volume-3, Issue-3, March 2017 Pages 52-55

Twitter Based Information Extraction

Manjeet Kumar, Abhishek Garg, Anuj Munjal, AkanshaTanwar

***Abstract*— In the modern world of social media dominance, the microblogs like Twitter and Facebook are probably the best source of up-to-date information. The amount of information available on these platforms is huge, although most of it is unstructured and redundant which makes our task of extracting information from it much more challenging. This automatic extraction of information from noisy sources has opened up new opportunities for querying and analyzing data.**

This paper is a review of the research that has been done on extracting information like event dates [1] and classification of information from social networking platforms like Twitter. We present a brief study of the work which shows that extracting useful information from Twitter and other social media platforms is indeed feasible. We provide brief study about the extraction techniques applied by the applications based on this subject like the extraction tasks and the input exploited for extraction, the types of methods of extraction used and the type of output produced.

***Index Terms*— information, extraction, twitter, extracting information from twitter, microblog.**

I. INTRODUCTION

Information extraction is the process of extracting useful information such as entities, relationships between entities from unstructured or raw data. Like from a random twitter post about a soccer match if we try to extract score and scorers of the match and store it structurally by classification of text. Information extraction, which is a sub topic of Natural Language Processing (NLP), now involves many different groups like Machine Learning and document analysis and has kept researchers from various parts of the world busy for almost three decades now [9].

a. Challenges

These social media platforms like Twitter and Facebook are probably one of the best and fastest ways of up-to-date news about ongoing activities around the globe. Almost every kind of news is of the interest of some or the other individual. This process of extraction of structure from such noisy sources like microblogs (like Twitter or Facebook)is indeed challenging. Moreover, given the fact that most of these posts are random tweets of daily activities like what some individual had for dinner or which concert they been to or even jokes and sarcastic posts, such events are only of immediate interest and hardly mean any value to the people apart from their own social circle and hence makes the task of retrieval of useful and targeted data far more challenging.

b. Scope

The self-centered nature and randomness of tweets are still a challenge to the state-of-art NLP butthe volume of Tweets is also much larger than the volume of other sources of news, so redundancy of tweets can be put to use more easily for getting the desired data. Millions of tweets are daily posted on twitter but most of them are either irrelevant orsuperfluous.

Applications like TwiCal[1] and TwitIE[2] are some of the work done on extracting events from Twitter and classifying text into domains like movies, sports, fashion etc. Because of the redundant nature of tweets, researchers tend to focus on extracting an aggregate representation of events which gives additional context for tasks such as event classification into domains, and also removesirrelevant events by exploiting redundancy of tweets (*Alan Ritter et al.*, 2011 [1]).

c.* *Input resources

The research work that has been discussed in this summary paper mostly uses tweets from the social networking platform Twitter as input resource for extraction of useful.

II. APPLICATIONS

The twitter based information extraction procedure is probably the most challenging task given the unstructured nature of the data present on these social networking platforms. Still a numbers of research papers have been published on this over the years.

Given the variety of content on these platforms, Natural Language processing methods can be applied to infer almost every type of information, sports, products, economy, etc. A few of the important work that we came across is listed as follows:

a. TwiCal

An application used to extract important event dates,*(Alan Ritteret al., 2011*[1]*).* This application is based on retrieving useful dates (e.g. product launch or an important soccer match) from twitter posts and then storing them into a calendar. It has been achieved using structured sequence-labelling to classify posts based on Verbs and Nouns(TwiCal-Events) and then extracting and resolving temporal expressions to know when these events are going to occur. Twical extracts a 4-tuple representation which includes entity, event, date and type. Extracted events are categorized into types based on variable models that infers set of event types to match data, and also classifies events into types by leveraging unlabelled data. It uses a method based on latent variable model inspired from work on modelling selectional preferences and unsupervised information extraction [1].

Manjeet Kumar,CSE,BharatiVidyapeeth's College of Engineering, New Delhi, India 110063
Abhishek Garg,CSE,BharatiVidyapeeth's College of Engineering, New Delhi, India 110063
Anuj Munjal,CSE,BharatiVidyapeeth's College of Engineering, New Delhi, India 110063
AkanshaTanwar, CSE Department, BharatiVidyapeeth's College of Engineering,New Delhi, India 110063

b. TwitIE

Its focus is on information extraction, opinion mining, summarization, visual analytics and user and community modelling. It is based on the GATEalgorithms given by (*Cunningham et al.*, 2002 [2]). Information is extracted from tweets by going through following steps: Language identification, tokenisation, sentence splitter, POS tagger, gazetteer lists, finite state transducer (based on [2] built-in regular expressions over annotations language), orthomatcher and coreference resolver. Data is delivered from the Twitter API in JSON format. Each tweet objects text value is changed into the document, which is covered with an annotation whose features represent all the key value pairs in the tweet JSON .The TwitIE system uses the *TextCat*(Cavnar and Trenkle, 1994) language identification algorithm, which relies on n-gram frequency models to discriminate between languages.

c. Keyphrase Extraction

It gives a way to extract topical keyphrases as a way to summarize twitter (*Zhao et al*., 2011 [3]). It uses a context-sensitive PageRank method for implementing keyword ranking and probabilistic function which takes both relevance and interestingness of keyphrases for performingkeyphrase ranking. It consists of three main steps of keyphrase extraction namely: ranking of keyword, keyphrase generation and then keyphrase ranking. A context sensitive topical Pagerank method (cPTR) for performing the first step of keyword ranking and probabilistic scoring function for third step of keyphrase ranking is being used.

d. Finding events

Main approach to locating events in social media as written by *Benson et al.*, 2011 [4], is to accurately induce some events from messages, evaluated against the events from a local city guide. It follows a structured graphical model which simultaneously analyses individual messages, clusters them according to event, and induces a canonical value for each event property.It biases local decisions made by the CRF to be consistent with canonical record values, thereby facilitating consistency within an event cluster. It employs a factor graph mode to capture the interaction between each of decisions. Variational inference techniques allows to effectively and efficiently make predictions on a large body of messages.The output of model consists of an event-based clustering of messages, where each cluster is represented by a single multi-field record with a canonical value chosen for each field.

e. Identification of entities and relationships

This extracts named entities from tweets. As proposed by Xiaohua Liu, Shaodian Zhang, Furu Wei, Ming Zhou [5],firstly, k-Nearest neighbours (KNN) based classifier is adopted to conduct word level classification and then fed to the linear conditional random fields (CRF) model which conducts fine grained tweets level NER. Furthermore, KNN and CRF model are repeatedly retained with an incrementally augmented training set, into which highly confident label tweets are added. Finally, gazetteers are used which cover common names, countries, locations, temporal expressions etc. NER task is divided into two subtasks i.e. boundary detection and type classification. As given by *Liu et al.*, for each word in the input tweet, a label is allocated to it indicating both the boundary and type of entity [5].

f. Classification of text(tweets):

The proposed approach by Bharat Sriram [6], effectively classifies given text into a predefined set of generic classes. Based on the number of classes present, there are two major types of classification: Binary – which classify input objects into one of the two classes. Multi-class – which classify input objects into one of the multiple classes. It proposes an intuitive approach to determine the labels of classes and the features with a focus on user intentions on Twitter such as conversations, sharing information/URLs. Next, it allows users to add new categories based on their interest. As the accuracy decreases with increase in number of classes used, it also permits users to add some new features relatedto the new classes.

III. TYPES OF STRUCTURES EXTRACTED

Unstructured data is the data that is not structured (data stored in various fields of a database, that is, it is not in the form of rows and columns. Unstructured data includes text and multimedia content such as emails, photos, videos and many other documents. These files may have an internal structure but they are still considered to be unstructured because the data they contain cannot be conveniently assimilated in a database.

The types of structure that can be extracted from an unstructured source can be categorized into four types: Entities, Relationships between Entities, adjectives describing Entities, and high order structures such as Lists and Tables.

a. Entities

Entities are real life objects about which the information can be stored. In unstructured data, entities can also be the various noun phrases. The most popular entities are named entities like names of people, places and organizations. Named entities consist of three subtasks: proper names and acronyms of persons, locations, and organizations

(ENAMEX),*N. A. Chinchor, 1998* [7], absolute temporal terms (TIMEX) and monetary and other numeric expressions (NUMEX). Now the term entity has modified to also include generic terms like diseases, proteins, paper titles, and journals. The ACEcompetition for entity relationship extraction from natural language text lists more than 100 different entity types.

b. Relationships

Relationships describe the way in which the instances of an entity are related to the entity of another entity. A relation can be unary, binary, tertiary and so on. The extraction of relationships is different from the extraction of entities as entities refer a sequence of words in the source whereas the relationships express the association between different snippets of texts representing different entities. Record extraction is the extraction of N-ary relationships. A popular subtype of record extraction is event extraction.

c. Adjectives Describing Entities

In many applications, there is a need to associate certain entities with values of adjectives to better describe the entities. The values of these adjectives are generally derived from the many words surrounding the entity. For example, we can do sentiment analysis of the entity, that is, we can infer whether the entity has a favourable or unfavourable opinion about the document. This is also called opinion extraction and is a very active topic for various research interests.

d. Structures such as Lists, Tables and Ontologies

The extraction systems have upgraded exponentially and now also include extraction of data from richer structure such as lists, tables, trees and ontologies from various documents instead of extraction just from plain text documents.

IV. METHODS OF EXTRACTION

The methods for extraction of information can be categorized along two dimensions: hand-coded or learning-based and rule-based or statistical.

a. Hand-coded or learning-based

Hand-coded system requires a programmer as well as a domain expert to define the rules and regulations or regular expressions or fragments of program for performing the extraction of information. The programmer should also possess proper understanding of linguistics so as to develop robust rules for extraction of information. On the other hand, the learning-based systems uses machine learning models for extracting information. These machine learning models are developed with the help of manually labelled examples of extracting information from an unstructured source. Domain knowledge is also required in learning-based systems to identify and label various examples that will represent how it will actually de deployed. Knowledge of machine learning is also of utmost importance so as to be able to choose from various alternative models and also so that the system can work properly for any new or unseen data. The noise in the unstructured data and the nature of the data will determine whether hand-coded system or learning-based system is preferred for the process of information extraction.

I. Rule-based or statistical

In rule-based systems a set of rules are predefined either manually or learned through the help of machine learning which are used for information extraction. The text, represented by features is compared to the rules and if there is a match the rule is fired up. Rules are made up of patterns and actions. The text tokens are compared to these patterns and if a match is found, the action is fired up. On the other hand, statistical model for extraction converts the extraction to be performed to a problem of designing and decomposing of the unstructured text and then to label the different decompositions. The labelling may either be independently that is the decompositions do not depend on other decompositions or it can be jointly if the decompositions are dependent on each other. The most common form of decomposition is the process of tokenization which is done with the help of breaking the unstructured text with the help of various predefined delimiters (like spaces, commas and full stops or dots). The various tokens are then assigned an entity label or an entity subpart label.

V. FUTURE SCOPE

In the recent times, microblogs have taken over news channels and daily news journals as our fastest source of information thanks, to the exponential growth in the number and variety of users with present on these platforms. And this impact of social media is only going to increase in the future which gives us more opportunities to explore and innovate in this domain. With better techniques to remove redundancies of tweets and also to filter them according to the need can lead us to amazing result given the amount of information available on Twitter. With continuous development in NLP techniques, Machine Learning and other related areas, information present on microblogs can be used to study things in deep, like sentiment analysis or even prediction analysis.

VI. CONCLUSION

This paper gives a review of some of the research work done on extracting information from social media platform Twitter and also of the information extraction state-of-art to some extent. This paper studies in brief some of the applications based on retrieving useful data from microblog. As already discussed in the paper, information extraction is a rather challenging task and given the amount of data available on social sites like Twitter can still be optimized to be

implemented at better level by using a better approach towards eliminating the unwanted tweets.

REFERENCES

[1] Alan Ritter, Mausam, Oren Etzioni, Sam Clark, "Open domain event extraction from twitter", University of Washington, 2011.

[2] H. Cunningham, D. Maynard, K. Bontcheva, V. Tablan. GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications.Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02). Philadelphia, July 2002.

[3] Xin ZHAO, Jing JIANG, Jing HE, Yang SONG, Palakorn ACHANANUPARP, 2011, "Topical Keyphrase Extraction from Twitter", proceedings of the 49th Annual Meeting of the Association for computer linguistics, Portland, Oregon, pp. 379-388, June 2011.

[4] Benson, A. Haghighi, and R. Barzilay. "Event discovery in social media feeds." In ACL, 2011.

[5] Xiaohua Liu, Shaodian Zhang, Furu Wei, Ming Zhou. "Recognizing named entities in tweets", Harbin Institute of Technology, Shanghai Jiao Tong University and Microsoft Asia.

[6] Bharat Sriram, "Short text Classification in Twitter to Improve Information Filtration", Ohio State University.

[7] N. A. Chinchor, Overview of MUC-7/MET-2, 1998.

[8] KalinaBontcheva, Leon Derczynski, Adam Funk, Mark A. Greenwood, Diana Maynard, NirajAswani "TwitIE: An open-source information extraction pipeline for microblog", University of Sheffield.

[9] Sunita Sarawagi, "Information Extraction" (survey), IIT Bombay and Microsoft Research.