Sociopedia: An Interactive System for Event Detection and Trend Analysis for Twitter Data

R. Kaushik, S. Apoorva Chandra, Dilip Mallya, J.N.V.K. Chaitanya and S. Sowmya Kamath

Abstract The emergence of social media has resulted in the generation of highly versatile and high volume data. Most web search engines return a set of links or web documents as a result of a query, without any interpretation of the results to identify relations in a social sense. In the work presented in this paper, we attempt to create a search engine for social media datastreams, that can interpret inherent relations within tweets, using an ontology built from the tweet dataset itself. The main aim is to analyze evolving social media trends and providing analytics regarding certain real world events, that being new product launches, in our case. Once the tweet dataset is pre-processed to extract relevant entities, Wiki data about these entities is also extracted. It is semantically parsed to retrieve relations between the entities and their properties. Further, we perform various experiments for event detection and trend analysis in terms of representative tweets, key entities and tweet volume, that also provide additional insight into the domain.

Keywords Social media analysis · Ontology · NLP · Semantics · Knowledge discovery

R. Kaushik (✉) · S. Apoorva Chandra · D. Mallya · J.N.V.K. Chaitanya ·
S. Sowmya Kamath
Department of Information Technology, National Institute of Technology Karnataka,
Surathkal, Mangalore 575025, India
e-mail: kaushik1603@gmail.com

S. Apoorva Chandra
e-mail: apoorvachandras@gmail.com

D. Mallya
e-mail: dmallya93@gmail.com

J.N.V.K. Chaitanya
e-mail: chaitanya.jnvk@gmail.com

S. Sowmya Kamath
e-mail: sowmyakamath@nitk.ac.in

A. Nagar et al. (eds.), *Proceedings of 3rd International Conference on Advanced Computing, Networking and Informatics*, Smart Innovation, Systems and Technologies 44, DOI 10.1007/978-81-322-2529-4_6

1 Introduction

Modern day social media are already a valuable source of information that exhibit the inherent qualities of volume, variety, veracity and velocity, making social media analysis a big data analytics problem. Corporations use social media to map out their demographics and analyze customer responses. People generally rely on the opinion of their peers on social media to shape their own opinions. Social media, then, can be considered to be a strong biasing factor in the era of Web 2.0 [1]. Various social networking sites also act as a source of information for its users and for analytics applications alike. Dissemination of news and trends has been proven to be faster on social media than by conventional media or even news websites. Twitter and Facebook tend to be the most commonly used Web 2.0 services on the Web and a big part of the fabric of web savvy population's daily life.

The widespread acceptance of social media as a source of useful data is based on the fact that, the interactions made by the users are of great social significance and are highly dependent on both the user's sentiments and also on peer consensus. Taking micro-blogging as an example, Twitter currently has over 280 million active users and over 500 million tweets are sent per day. Using such a high volume and velocity data to provide useful information has become a tedious task bordering on the impossible. Automating this process through intelligent, semantic-based information extraction and knowledge discovery methods is therefore increasingly needed [2]. This area of research merges methods from several fields, in addition to semantic web technologies, for example, Natural Language Processing (NLP), behavioral and social sciences, data mining, machine learning, personalization, and information retrieval [3].

The problem of making sense of the data we get from various websites like Twitter is currently an area of huge research interest. It is near impossible to perform manually and we need to utilize various big data and semantic web techniques to gather usable information from the Twitter feed. The proposed solution discussed in this paper, relies on the concept of an ontology to understand the relation between various terms used in the tweets and for enabling automatic analysis of the data. The ontology is constructed from the tweet dataset and the relations between terms are mapped. We then use the constructed ontology to provide statistics and usable information from the data to serve an user inputted search query.

The rest of the paper is organized as follows—In Sect. 2, we discuss various related work in the literature and also some relevant ontologies that currently exist for social media representation. Section 3 discusses the proposed methodology and the various components of system and the development process. In Sect. 4, we present experimental results observed and discussion, followed by conclusion and future work.

2 Related Work

The different kinds of social media coupled with their complex characteristics, make semantic interpretation extremely challenging. State-of-the-art automatic semantic annotation, browsing, and search algorithms have been developed primarily on news articles and other carefully written, long web content. In contrast, most social media streams (e.g. tweets, Facebook messages) are strongly inter-related (due to the streams being expressions of individual thought and experiences), highly temporal, noisy, bursty, short, and full of colloquial references and slang, thus making it difficult to generate accurate analysis, without semantic and linguistic parsing.

Iwanaga et al. [4] presented a ontology based system for crisis management during crucial time periods like during an earthquake. The requirement for an ontology in that scenario is to ensure the proper evacuation of the victims. Their dataset consisted of tweets on the earthquake in Japan's Tohoku region in Japan. Several researchers have concentrated on the issue of learning semantic relationships between entities in Twitter. Celik et al. [5] discuss the process of inferring relationships between the various entities present in a tweet such as timestamp, unique id and # tags and @ tags. There exist many methods for inferring relationships [6, 7] such as using already existing ontologies and web documents crawling, using a bag of words approach based on the term frequency and inverse document frequency of the terms involved and using term co-occurrence based strategies. They concluded that co-occurrences based approaches result in the highest precision.

Bottari et al. [8] proposed an ontology, which has been developed specifically to model relationships in Twitter, especially linking tweets, locations, and user sentiment (positive, negative, neutral), as extensions to the Socially-Interlinked Online Communities (SOIC) ontology. Unlike SIOCT, Bottari identifies the differences between retweets and replies. Geo-locations (points-of-interest) are represented using the W3C Geo vocabulary, and hence can be used to develop location-based logic and relations. For interlinking Social Media, Social Networks, and Online Sharing Practices, DLPO (LivePost Ontology) is used, which provides a comprehensive model of social media posts, going beyond Twitter. It is strongly grounded in fundamental ontologies, such as FOAF, SOIC, and the Simple Knowledge Organization System (SKOS). DLPO models personal and social knowledge discovered from social media, as well as linking posts across personal social networks. For modeling tag semantics, Meaning-Of-A-Tag (MOAT) ontology can be used allowing users to define the semantic meaning of a tag by linking Open Data and thus, manually creating semantic annotations within the social media data itself. The ontology defines two types of tags: global (i.e. across all content) and local (referring to a particular tag on a given resource). MOAT can be combined with SIOCT to tag microblog posts.

3 System Development

3.1 Data Cleaning

Figure 1 depicts the various processes of the proposed system. Since it is intended for the use of company officials for gathering the social media users' opinion about their brands, we need social media data, for which we chose Twitter. Tweets containing particular keywords were extracted from Twitter using a tool called Sysomos [9], which also provides the URL, associated timestamp, location, sentiment and the user who tweeted. We obtained as average of about 40,000 tweets on each of the datasets—namely Apple's *iphone6*, Chinese Manufacturer, *Xiaomi's* smartphones and Motorola's *Moto* smartphone.

As the first step to cleaning, we filtered the tweets to retain only English language tweets, as natural language processing is supported for English only, currently. This was done by analyzing the stopwords found in the tweet and assigning the tweet a ratio for each language that would indicate its chance of belonging to that language. After the ratio was assigned, the tweets in English obtained the maximum scores and all tweets with scores above a threshold were selected. Tweets with too many # and @ are considered to be spammy tweets and are automatically eliminated during stopword detection technique. The remaining English tweets are collected along with timestamp and stored.

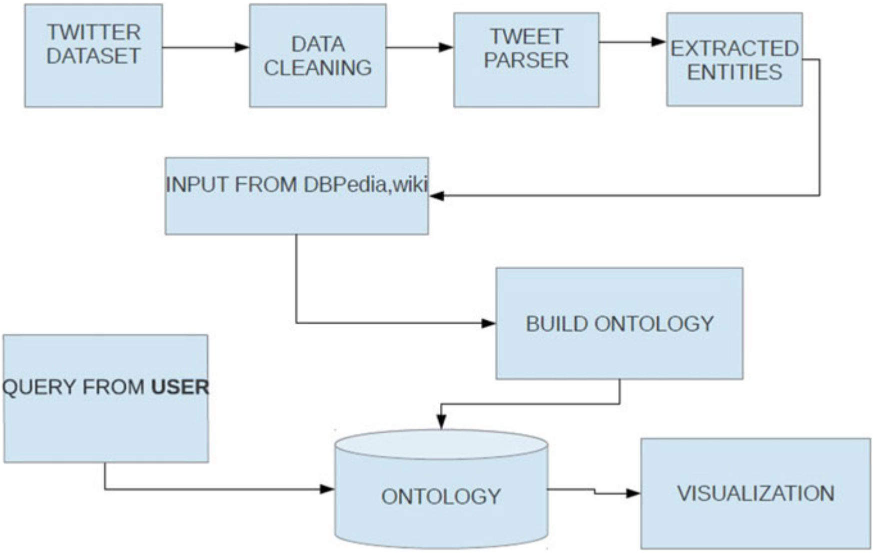

Fig. 1 Workflow of the proposed system

3.2 Data Preprocessing

After tweet cleaning, tweets are sorted according to timestamp and are counted on a daily basis. The primary requirement of this application is that the company executives must know about the current social media buzz about their products. This can be best indicated by computing the volume of tweets generated during the day and plotting the volume versus time. The granularity level was maintained as a day because lesser than that would result in the appearance of many peaks which would be difficult for the marketer to analyze and many of them may not be interesting events as well.

The system as to provide a search mechanism that can help the marketer to research and analyze the topics that were being discussed in the social media. In order to build such a search engine, we used an novel approach using an ontology constructed from the tweets to capture the trends and relationships between events. An ontology is a collection of entities and relationships between them. The keywords or entities extracted from the tweets are to be represented in the constructed ontology. These entities or keywords were extracted from the tweets using the CMU tweet parser [10, 11].

The CMU tweet parser tokenizes tweets and returns the parts of speech tagging for each of the tweets that are given as an input. The confidence associated with each part of speech was also returned after the tagging [12]. The output of the parser was stored in a comma separated file. A separate parser was required for this purpose because a POS tagger that works on formal sentences cannot be used with tweets, where many local abbreviations are used due to the limitation in tweets (only 140 characters can be present at the maximum in a tweet).

After POS tagging, the entities that were used in the further analysis are *named entities, hashtags* and *nouns* as these conveyed the maximum information about the topic. Figure 2 shows the percentage of each in the resultant processed dataset. These entities were extracted from the tweets and a frequency distribution was generated for all extracted entities. From this frequency distribution, the top 100 keywords were chosen as final entities based on which the ontology would be built. The frequency of occurrence of these entities were also stored.

After selecting the top 100 keywords as entities for constructing the ontology, the next phase is to capture any relationships between these entities. The ontology can be complete only if relationships exists between the entities. Since, the tweet data retrieved from Twitter cannot be considered as a reliable published source of information in the standard sense, it is imperative to establish relationships between documents using other techniques. We chose Wikipedia as a source of reliable published information, available as a structured dataset, to recognize the relationships between entities. We systematically retrieved documents from Wikipedia, which are then stored in a directory, for later access during the ontology building process.

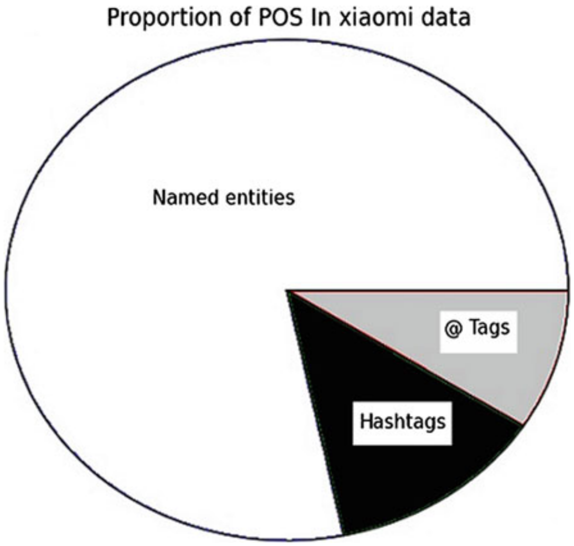

Fig. 2 Entities extracted using CMU Tweet Parser

3.3 Build Ontology Module

The ontology is a graph or a network of nodes (entities) that are connected by edges (relationships identified by relationship names). The relationships between the extracted entities are inferred using Wikipedia, DBpedia and web documents for the extracted entities. The ontology is built using Web Ontology Language (OWL), as it is a standardized ontology language based on description logic. The generated ontology is saved in an ontology base which is one of the main underlying components for the search engine to be built.

Currently, we provided a basic search interface, that works using the constructed ontology saved in the ontology base. The user can enter a query which will be processed to retrieve documents regarding the topic on which the ontology was developed, corresponding to most recent events and trends. The results of the search engine will depend upon the type of the dataset.

4 Results and Analysis

We considered two different tweet datasets pertaining to the launch of two popular products, Motorola Moto series of mobile phones and Apple's iPhone 6. The *Moto Dataset* contains 40,000 tweets collected during the period Sep 1st to Sep 9th 2014 and the *iPhone6 dataset* contains about 35,000 tweets collected during the period Sep 6th to Sep 12th 2014. As seen from Fig. 3, spikes were detected in the tweet volume versus time graph on Sep 2nd to 5th, i.e., during the time when Moto

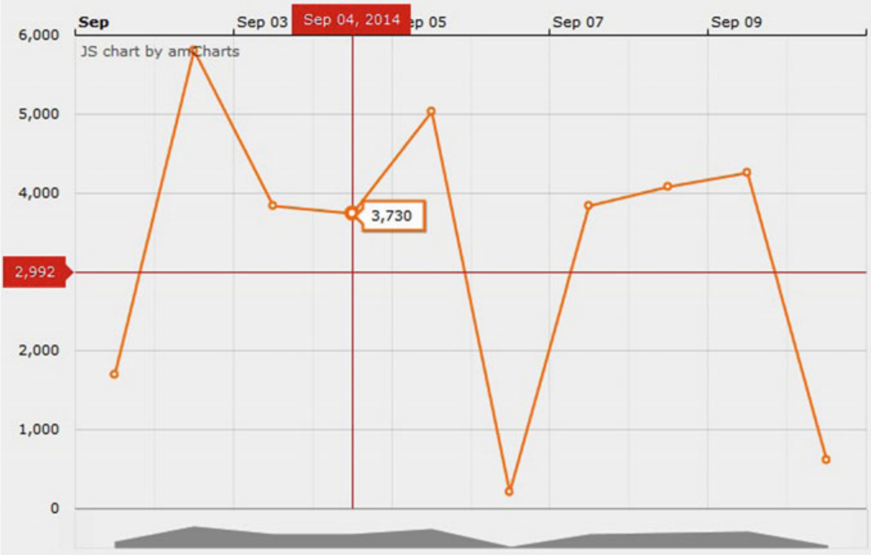

Fig. 3 Tweet volume versus time for the Motorola Dataset

released the next generation of the Moto X and the Moto G smartphones and Moto 360, a wearable smartwatch Android device. This is an indication that the system was able to detect events that pertain to certain events interesting to the target user, that is release of a product and the hype surrounding it during the initial days.

Certain other interesting observations could be made from the work that was carried out. Out of all the entities generated by the CMU tweet parser, those corresponding to the nouns, hashtags and the named entities were the most useful. Verbs, even though may be useful in establishing relationships between the entities can be obtained from Wikipedia and other the web documents that were used to infer relationships. Hashtags generally form a lesser proportion of the entities after parsing the tweets, so some weightage can be given to hashtags, as information gain from hashtags can be significant.

5 Conclusion and Future Work

In this paper, we proposed an interactive system for event detection and trend analysis of Twitter data. It is intended to automatically detect trends in social media regarding certain real world events like launch of new products, advertising campaigns etc., to aid companies to analyze the effectiveness of their marketing approach or the mood of the consumer through their tweets on Twitter. Currently, available ontologies are mostly based on web documents or direct information from sources as a result of which, they cannot be contemporary to any event/topic they are related to [13]. In contrast, we are attempting to build an ontology from social media as a result of which the ontology will be contemporary to currently occurring events, a major advantage for building a search engine, whose results change dynamically with time.

References

1. Asur, S., et al.: Predicting the future with social media. In: 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), vol. 1 (2010)
2. Ritter, A., et al.: Unsupervised modeling of twitter conversations (2010)
3. Mika, P.: Flink: semantic web technology for the extraction and analysis of social networks. Web Semantics: Sci. Services Agents World Wide Web **3**(2), 211–223 (2005)
4. Iwanaga, I., et al.: Building an earthquake evacuation ontology from twitter. In: 2011 IEEE International Conference on Granular Computing (GrC) (2011)
5. Celik, I., et al.: Learning semantic relationships between entities in twitter. In: Web Engineering, pp. 167–181. Springer (2011)
6. Ozdikis, O., et al.: Semantic expansion of hashtags for enhanced event detection in twitter. In: 1st International Workshop on Online Social Systems (2012)
7. Owoputi, O., et al.: Improved part-of-speech tagging for online conversational text with word clusters. In: HLT-NAACL, pp. 380–390 (2013)
8. Celino, I., et al.: Towards bottari: using stream reasoning to make sense of location-based micro-posts. In: The Semantic Web: ESWC 2011 Workshops, pp. 80–87. Springer (2012)
9. Sysomos.com: Sysomos heartbeat
10. Gimpel, K., et al.: Part-of-speech tagging for twitter: annotation, features, and experiments. In: 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers, vol. 2, pp. 42–47 (2011)
11. Kong, L., et al.: A dependency parser for tweets. In: International Conference on Empirical Methods in Natural Language Processing, Doha, Qatar (2014)
12. Derczynski, L., et al.: Twitter part-of-speech tagging for all: Overcoming sparse and noisy data. In: RANLP, pp. 198–206 (2013)
13. Zavitsanos, E., et al.: Gold standard evaluation of ontology learning methods through ontology transformation and alignment. IEEE Trans. Knowl. Data Eng. **23**(11), 1635–1648 (2011)