International Journal of Engineering & Technology, 7 (4.19) (2018) 880-883

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Twitter Information Representation using Resource Description Framework

Asaad Sabah Hadi*[1], Enass Mahdi Abdulshaheed[2]

[1,2]Babylon, University of Babylon, Iraq
**Corresponding Author E-mail: asaadsabah@yahoo.com*

Abstract

Social media service emerged as essential forums for negotiating and giving comments about news, events happening around the world. Such user-generated information can be important data source for researches.Twitter is a micro-blogging which is growing rapidly in the last years.Due to the fast evolution of twitter, the Researchers study tweets' content characteristics, so this can help in extraction of information like users' information, opinions about topics and other useful information from tweets.This work proposed a method to extract useful data from tweet object and represent it in Resource description framework (RDF) data model. The Streaming API used to collect twitter data by filtering the results based on specific keywords. The response obtained as Java Script Object Notation (JSON) file,then while parsing the obtained file thetweet's text preprocessing and extract useful information from other entities of tweet object, Jena API used to represent extracted data in RDF data model and save it in RDF file to be used in future applications.

Keywords: *Social Media, Twitter, Tweets, RDF, Jena, Information Extraction.*

1. Introduction

Social media services emerged as essential forums for negotiating and giving comments about news, events happening around the world. Such user generated information can be important data source for researches in different fields such as data science, sociology, psychology and historical studies that can be used by researchers to understand behavior, trends and opinions [1]. Twitter is a micro-blogging which is growing rapidly in the last years .it is considered as online social network service that counts about 336 million monthly active users at the beginning of 2018 [2].

Twitter enable users to send and read posts that are limited to 140 characters for each post (tweet).People can communicate and share their opinions about products or movies and upcoming events such as sports or political elections, etc. [3]. Due to the fast evolution of twitter, the Researchers study tweets' content characteristics, so this can help in extraction of information like users' topics, opinions about topics[4][5].Tweet contains important information but there is a noise as result of tweet's shortness, marks, irregular words,noise, relevance, emoticons, folksonomies and slangs,etc. [6][7].

The contribution of this paper isto extract useful data from tweet object in addition to extract important entities from tweet's text and represent it in RDF data model to be used in future in advanced steps of the proposal. After retrieving the tweets' JSON file by using Twitter streaming API, preprocessing text and data extraction from entities of tweet object. The Jena API used to create RDF data model and save extracted data in RDF file. The file willbe used in future step of the proposed system or future applications. The work flow diagram presented in proposed system section.

2-Social Media

Social media considered as internet application. It can be categorized as web-based and mobile based applications. It facilitates the creation, access and exchange of data generated by the users ubiquitously. Social media provide the opportunities to understand the behavior of individuals and society. Business, bioscience, social science can be considered the application of social network. Data can be gotten from social media by programmable API or Tools. Social media data formats like XML and JSON that can be financial data, customer transaction data, telecoms and spatial data[9]. Social media archives emerged as source of researches in multiple fields, like data science, sociology or the digital humanities, so Twitter is considered as most prominent source.[1]

3. Twitter

Twitter provides message services that count about 336 million monthly active users at the beginning of 2018[2], user can send or receive massage no more than 140 characters called (tweet), this limitation can be considered important property to increase and facilitate the post, share and forward operations. There is no need to user permission to be followed or to be a follower of other user. Twitterers can share their ideas, feelings, trends, media, news, and opinions. Companies, Celebrities and politicians can be followed by twitter's users to evaluate products and behaviors. Also the use of Hashtags facilitates searching of tweet topics[10][4][3][5].

Tweets contain entities and places in addition of text information. Entities like user Hashtags, mentions, URLs associated with a tweet, places represent the locations in the real word so tweets play essential role in twitter data analysis[11].

Tweets can be available to researchers through free APIs that can be classified into Search, Restand Streaming APIs, these API can be accessed by authenticated requests. The fetched data by these APIs in JavaScript Object Notation (JSON) format[8][12][9]. Tweets are saved in JSON file as one array of objects per line for each one. each object contains key value pairs for each attribute and its value. In addition to text, tweet'sobject involves additional meta data like user and entities [14].

4. Linked Open Data

Linked data essential goal is to structure and interconnect data by semantic web technologies, like Resource Description Framework (RDF)[13]. Data surround us everywhere and every time to describe the performance, so data can play an essential role in our life to make the best decisions. So individuals, organizations, governments and researchers need to share data. This data need to be in consumption by Third parties, to build new businesses, online commerce, fast scientific progress, and enhancement of the democratic process.Due to development and variety of ecosystems, there is a need to access these data. World Wide Web (WWW) has changed the manner we connect and consume documents, so can changed the way we discover access, integrate and use data. The Web can be the best medium to enable these processes, because of its ubiquity, its distributed and scalable nature, and it's mature. The linked data depend on the mechanisms of reusing and sharing data. The well-structured data enable the easily creation of tools to process it for reuse. The structured data are available on the Web by Web APIs. Web APIs supply simple query access to structured data over the HTTP protocol [15].

Structured data are provided by Web APIs in XML and JSON formats, which have supported by programming languages. Web APIs provide the access of data on the Web and not place it truly in the Web, let it linkable and discoverable.Linking data is distributed across the Web by a standard mechanism to specify the existence and meaning of connections between items described in this data. Resource Description Framework (RDF) provides this mechanism. So, Linked Data can be defined as set of best practices for publishing and interlinking structured data on the Web. These practices are known as principles of linked data and must be applied [16].The principles as following: [15][16]

1. Using URIs as names for things;
2. Using HTTP URIs, so these names can be looked up;
3. Someone can look up a URI to provide useful information, using the standards (RDF, SPARQL);
4. Links can be included to other URIs to discover more things.

5. Semantic Web

The interpretation of the huge data in theWorld Wide Web (www) done by humans and there is no role for machine. While in semantic web data interpretation can be done by machine to support the user in his task[17]. In Semantic web Search would not be as looking for keywords, but looking for synonyms and being aware of homonyms, and consider context and purpose of the search query. Agents have the ability to understand the contents of a web page and tailor it to personal interest profiles [16].

Things are referred as resources In the Semantic Web; a resource may be anything that might talk about. Shakespeare, Stratford, "the X value," and "all the cows in Texas", all these examples of things and can be someone talk about and considered as resources in the Semantic Web. Resource is the Semantic Web standard that represents the name of the base technology in the Semantic Web Resource Description Framework (RDF)[18].

Web content can be processed by wide range of applications by agree on standardized content formats. So, when publishing linked data on the web it must be represented in RDF.RDF is a data model that describes a resource as a number of triples. Each triple has three parts (Subject, Predicate, and Object) as shown in Table1.[15][16].

Table 1: RDF Triple

Subject	Predicate	Object
Matt Briggs	Has_nick_name	Matty

6. Proposed System

The work flow steps of the proposed system are showed inFigure (1) and described as follow:

6.1 Data Collection

The proposed system used Twitter Streaming API which it is a tool that enable the interaction with computer programs and web services. Pythonlibrary called **Tweepy** usedto access Streaming API to download real-time stream of Tweets, The first step of getting twitter data is to create twitter application to enable the access to Twitter API.The user must logging in his twitter account and opening the application sitehttps://apps.twitter.com/to create an application which can provide the authentication keys like access_token,access_token_secret, consumer_key and consumer_secret. The obtained keys are used in programming application to get twitter data depending on specified keywords, then saving the data in JSON file.

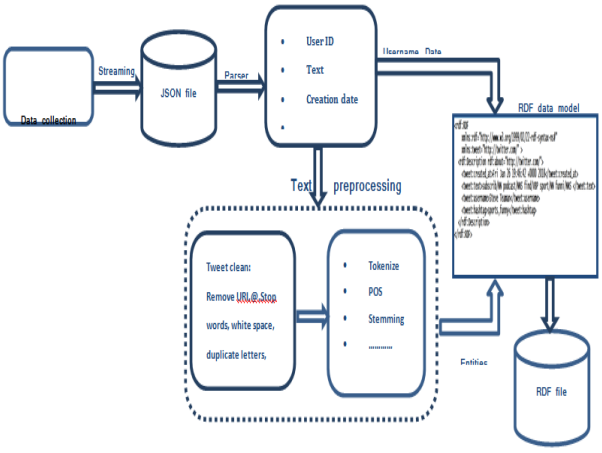

Fig. 1: work flow diagram for the proposed system

6-2 Parsing JSON File

The proposed system used Java programming language for parsing JSON file. It provides professional API to deal with JSON data format. Parsing operation done on English tweets only. Then extract specificinformationfrom tweet object like user name, date, Hashtags, Etc.

6.3 Tweet Text Preprocessing

This stage done in twosteps:

6.3.1 Tweet Text Cleaning

Tweet text is very noisywith linguistic errors and idiosyncratic stylebecause the users use different informal words and make misspells, repetition in characters and white spaces.Also text may have symbols like URI, RT, #, @,*, /,?&, %, $, etc. So this stage remove all symbols and duplicates and convert the capital letters to small letters also remove stop words (to, our, and…).

6.3.2 Entity Extraction From Text

In this step multiple operations are done on text result from previous stepto extract important entities as in thefollowing steps:

- Tokenize the text into individual words.
- Stemming of each word using Porter Stemmer API in Java language to remove all suffixes from words (ing, s, tion …).
- Part Of Speech (POS) for each word using Ark-tweet NLP API.Part-of-speech tagging considered as central problem in natural language processing and a key step to specify part of speech of each word in text.
- Entity Extraction from text using Gate PIA which provide specific Twitter tool (TwitIE)which is used for named entity recognition application.

6.4 RDF Model Creation

Java programming language provides JENA API to create and manipulate RDF data model. The proposed system considered the tweet as a resource and the extracted entities and their values from previous step considered as the properties (predicates) and its values as objects of the resource (subject). Then after the creation of the model it saved in (.rdf) file.All mentioned steps applied repeatedly on all tweets in JSON file to be represented in RDF data format.

7. Implementation

7.1 Tweet Data Collection

Data collection had done bylogging in twitter user account, and going to the application sitehttps://apps.twitter.com/which enable the researcher to create an application. It provides specific and unique keys:access-token, access-token-secret, consumer-key, and consumer-secret.The keys used in Python program application that access Twitter Streaming API fordownloading tweet data. Download operation had done on specified keywords (sports, computer science, breast cancer, Barcelona and Real Madreid, cars) in 26/Jan/2018, then saving the data in JSON file as in Figure(2), the size of downloaded file is 1 MB.

Fig. 2: Tweet JSON file

7.2 Parsing JSON File

Parsing operation had done using java programming language.Gson.jar API was used for the parsing operation. Parsing operation made on English tweets only. The manipulation of input data file can be on any file size but the used file size is 1 MB as a prototype of work. It contains 170 tweets in different languages; parsing operation had done for 134 English tweets. While parsing the JSON file, Information Extraction operations had done for each tweet object. The extracted information is user name, friends count, etc. from "user" object.Date extracted from "created at" value part and Hashtags from "entities" object if it is available, etc. as shown in Table 2.

Table 2: Information Extraction from tweet‘s object

Before extraction	After extraction	
"created_at":"Fri Jan 26 19:46:47 +0000 2018"	“Date”	26-Jan-2018
"user":{"id":3063856888,"id_str":"3063856888","name":"Steve Teaman","screen_name":"SteveTheC4veman","location":"570 scranton","url":null,"description":"22 panera bread is were I reside #570#wifedup #cowboynation #MGK #Laceup #heatnation #FloridaGators. how bout them cowboys!","translator_type":"none","protected":false,"verified":false,"followers_count":82,"friends_count":54,"listed_count":116,"favourites_count":8218,"statuses_count":8323……"profile_image_url_https":"https://pbs.twimg.com/profile_images/594599537848619009/OPReGDVg_normal.jpg","profile_banner_url":"https://pbs.twimg.com/profile_banners/3063856888/1430598550","default_profile":true,"default_profile_image":false,"following":null,"follow_request_sent":null,"notifications":null}	"name” "followers_count" “friends_count”	"Steve Teaman" 82 54

7.3 Tweet Text Preprocessing

This stagehad done in twosteps:

7.3.1 Tweet Text Cleaning

Cleaning operation of each tweet text had doneas in followingsteps which are shown in Table 3.

- Remove repetition in characters
- Remove white spaces.
- Remove URI.
- Remove symbols like RT,*, /?&, %, $, etc.
- Remove symbols like semicolon (;), column (,), dots (.), quotations (‘/”), etc.
- Remove arcs (, [. {,},]).
- Remove numbers.
- Remove stop words (to, our, and…).
- Convert capital letters to small letters.

Table 3: Cleaning tweet text

Noisy text	“text”:”RT @pardonmy_cte: Subscribe to our Podcast and find us at https://t.co/NvJUNI9B3zUniversity of California #sports #funny https://t.co/ZLPsFVwkii”
Clean text	“subscribe podcast find #sport #funnyUniversity California “

7.3.2 Entity Extraction from Text

In this step multiple operations are done on text result from previous step to extract important entities as in the following steps:

- Tokenize the text in to individual words.

- Stemming of each word using Porter Stemmer API in Java language to remove all suffixes from words (ing, s, tion …) ,such as:

Sports = sport

- Implement POS for each word using Ark-tweet NLP API [19] which it is Part-of-speech tagging considered as central problem in natural language processing and a key step to specify part of speech of each word in text such as shown in Table4.

Table 4: POS Tagging

Word	Subscribe	: -) (☺)	Sport	podcast	#sport
Tag	V	E	N	N	#

V for verb, E for emoticon, N for noun, O for object

- Entity extraction from text using Gate API which provide specific Twitter tool (TwitIE) which is used for named entity recognition application, such as in Table5.

Table 5: Entity extraction

Token	Entity
University	Organization
California	Place

7.4 RDF Model Creation

JENA API had used to create and manipulate RDF data model,so the extracted entities and their values from previous step considered as the properties (predicates) and its values as objects of the resource (subject) that represented by Tweet. Then the model saved in (.rdf) file as in Figure3.

Fig. 3: RDF file

The RDF validator in http://www.w3.org/RDF/Validator used to check the correctness of an RDF document. The result in Figure 4 showed the subject, predicate and objects for each entity in the tweet object and the model graph.

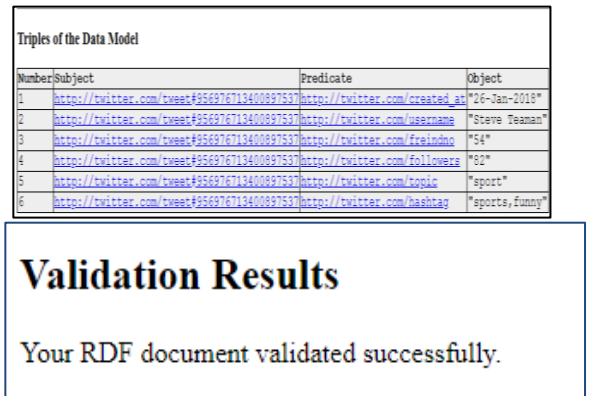

Fig. 4: the results of RDF validator

8. Conclusion

To post a tweetthere is no any restriction to give any idea. There is other limitations like:limited text length, most of people post false, incorrect information about events, spellings and grammar's error, the use of improper sentence structure and different language,thereforeit is difficult to distinguish important data from unused data. The main advantage of representation of tweets information in RDF format is making semantic meaning of tweets information rather than text processing of a whole tweets and it reduce the time because we focus on specific features rather than we process all information in collected data.

References

[1] Fafalios, F., Iosifidis, V., Ntoutsi, E. and Dietze, S. (2018). TweetsKB: A Public and Large-Scale RDF Corpus of Annotated Tweets. (ESWC'18), L3S Research Center.

[2] Statista – The portal for statistics, https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/, [Accessed 23 Jun. 2018].

[3] Bouazizi, M. and Ohtsuki, T. (2017). A Pattern-Based Approach for Multi-Class Sentiment Analysis in Twitter. IEEE Access, 5, pp.20617-20639.

[4] Kamel, N. (2012). Ontology-Based Information Extraction fromTwitter, 17-22.

[5] Asfari, O., Hannachi, L., Bentayeb, F., &Boussaid, O. (2013). Ontological Topic Modeling to Extract Twitter Users' Topics of Interest, ICITA , 141—146.

[6] Singh, T. and Kumari, M. (2016). Role of Text Pre-processing in Twitter Sentiment Analysis, Elsevier B.V., Volume 89, P. 549-554.

[7] Bao Y., Quan C., Wang L., Ren F. (2014) The Role of Pre-processing in Twitter Sentiment Analysis. In: Huang DS., Jo KH., Wang L. (eds) Intelligent Computing Methodologies. ICIC 2014. Lecture Notes in Computer Science, vol 8589. Springer, Cham.

[8] Goonetilleke, O., Sellis, T., Zhang, X. and Sathe, S. (2014). Twitter analytics. ACM SIGKDD Explorations Newsletter, 16(1), pp.11-20.

[9] Batrinca, B. and Treleaven, P. (2015). Social media analytics: a survey of techniques, tools and platforms. AI & SOCIETY, 30(1), p. 89-116.

[10] O'Reilly, T. and Milstein, S. (2012). The Twitter book. 2nd ed. Sebastopol, CA: O'Reilly, p.258.

[11] Russell, M. (2013). Mining the Social Web Data Mining Facebook, Twitter, LinkedIn, Google+, GitHub, and more, 2nd ed. O'Reilly Media, p. 448.

[12] Kumar, S., Morstatter, F. and Liu, H. (2014.). Twitter Data Analytics. 1st ed.Springer Briefs in Computer Science, Springer, p.77

[13] Vigo, M., Bellahsene, Z., Ienco, D. and Todorov, K. (2015). Twitter Event Detection and Modeling with TEWS. ISWC: International Semantic Web Conference, [online] Hal.archives-ouvertes.fr. Available at: https://hal.archives-ouvertes.fr/hal-01274024/ [Accessed 2 Jul. 2018].

[14] https://developer.twitter.com/en/docs/tweets/data-dictionary/overview/ intro-to-tweet-json.html/, [Accessed 25 Jun. 2018].

[15] Heath, T. and Bizer, C. (2011). Linked Data: Evolving the Web into a Global Data Space. 1st ed. Morgan & Claypool, pp.1-136.

[16] Antoniou, G., Van Harmelen, F., Hoekstra, R., Groth, P. and Antoniou, G. (2012). A semantic web primer. 3rd ed. Cambridge: MIT Press.p.270.

[17] Patel, G. A., &Madia, N. (2016). A Survey: Ontology Based Information Retrieval For Sentiment Analysis. IJSRSET 2(2), 460-465.

[18] Allemang, D., &Hendler, J. A. (2012). Semantic Web for the working Ontologist: Effective modeling in RDFS and OWL.2nd ed. Waltham, MA: Morgan Kaufmann/Elsevier.p.384.

[19] Gimpel, K., Schneider, N., Oconnor, B., Das, D., Mills, D., Eisenstein, J., . . . Smith, N. A. (2010). Part-of-Speech Tagging for Twitter: Annotation, Features, and Experiments. doi:10.21236/ada547371.