International Journal of Computational Intelligence Research
ISSN 0973-1873 Volume 13, Number 4 (2017), pp. 621-630
© Research India Publications
http://www.ripublication.com

Extracting Information from Social Network using NLP

Charu Virmani
Research Scholar, YMCAUST, India.

Dr. Anuradha Pillai
Ymcaust, Faridabad, India.

Dr. Dimple Juneja
NIT, Kurukshetra, India.

Abstract

While the popularity of Social Network is raising the field of Social network Analysis has become an important and interesting study in the area. Social Network analysis refers to the process of exploring social structures through the use of network and graphs. The information on the social network is unstructured and there is a need to extract the structured information for making use of the valuable information. Extracting information from the social network is the exploration that empowers the use of such a massive amount of unstructured distributed information in a structured way. Natural language processing is employed to enhance the accuracy in visualizing the structured information that is speckled over the social network. The foremost notion of monitoring is to analyze the meaningful information from texts written by naïve users of social network. It analyzes natural language text in order to extract information about different types of entities, relationships or events. The Natural Language methods are being looked closely by means of this research. In this paper researcher attempts to review various text mining systems which is the keystone of Natural Language Processing to analyze social network information.

Keywords: Natural Language Processing, Social Network, Information Extraction, Text Mining, Entity Extraction.

1. INTRODUCTION

There have been innumerable changes in traditional business methods that have been introduced post the arrival of social network. In recent times everything is conducted online beginning from development of the product to its marketing. It has become extremely easy and common that even identity of the customer is not revealed. This further makes the customers opinionated about a particular product they have purchased online, there are many business organizations which interact with the customers to know whether they are satisfied from the product or not. Social Network is applicable almost in every field, to name a few; we have, voting mechanism for beauty pageant, political campaigns, product research & promotion via advertisements. There is an arising need for analyzing and modeling of such networks. One can also inquire more about an organization's external environment with the help of competitive intelligence. Technology area is growing at a very rapid pace leading to formation of new sophisticated tools for text terms. Data mining techniques are required for their capability of handling the three dominant properties with social network data namely; size, noise and dynamism. This huge amount data of social network require automation for dispensation of information, analyzing it within a stipulated time. Interestingly, data mining techniques are designed to handle the voluminous data sets to mine significant patterns from data; social network sites provide these huge data sets because of their usage and hence are ideal candidates to mine data using the data mining tools. Therefore we can infer that the data mining or to be precise web mining provides the necessary intelligence to the social network to create and interact in a more humanly and user friendly manner. This paper is divided into five sections: section 1 provides the introduction, section 2 discusses the various techniques of Natural Language Processing, Section 3 discusses its challenges when imposed on social network and Section 4 provides the text mining approaches and section 5 concludes the paper.

2. NATURAL LANGUAGE PROCESSING (NLP)

This paper analyses excessive use of Natural Language Processing and web mining techniques to study Social Network. NLP techniques maps human language to machine language, it models the way user requests information to how computer or software understands it. However, simply searching for keywords is not an appropriate method in Social Network communication. Therefore one can observe that the encounter in Social Network monitoring is to extract and interpret ‘User Communication’. Some Network monitoring systems also uses NLP methods with statistical techniques to ensure the extracted information to be correct and precise. This paper discusses NLP approaches that are essential for Social Network monitoring which are Automatic Summarization, Chunking, Part-of-speech tagging, Named Entity Recognition, Named Entity Disambiguation, Fact/Relation Extraction, Word-sense Disambiguation and Sentiment Analysis.

A) Automatic Summarization

Automatic Summarization is the process of reducing a text document with the help of a computer program in order to create a summary that retains the most significant points of the original document. Technologies that can make a coherent summary take into account variables such as length, writing style and syntax. The main notion of summarization is to find a representative subset of the data, which contains the information of the entire set.

Generally, there are two approaches to automatic summarization: Extraction and Abstraction. Extraction refers to selecting a subset of existing words, phrases, or sentences in the original text to form the summary. In contrast, abstraction builds an internal semantic representation and then use natural language generation techniques to create a summary that is closer to what a human might generate. Automatic Summarization system takes three basic steps namely, Analysis, Transformation and Realization that are briefly explained below [1]:

In analysis, a concise and fluent summary of the most significant information is produced in the input. It requires the capability to reorganize, modify and merge information expressed in different sentences in the input. Transformation is an ordered text is generated by manipulating the internal representation post analysis in Auto Summarization. An analyzed summary text is generated using scores of transformation in the Realization phase. The process of Auto Summarization is depicted in Figure 1.

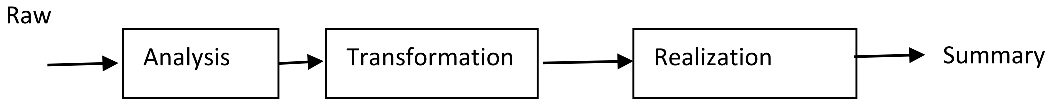

Figure 1. Process of Auto Summarization

B) Chunking

Chunking is the basic technique used for entity detection. Chunking selects a subset of the tokens rather than tokenization that omits whitespaces. The pieces formed in the source text do not overlap as the output of tokenization. It is easier to describe what is to be excluded from a chunk. It basically segments the tokens. A chink can be defined as a sequence of tokens that is not in a chunk. Removing a sequence of tokens from a chunk is refereed as Chinking.

The whole chunk is removed if the matching sequence of tokens spans an entire chunk. However, the tokens are removed, leaving two chunks where there was only one before; if it appears in the middle of the chunk. A smaller subset of chunk remains, if the sequence is at the periphery of the chunk.

The process of Text Mining is depicted in Figure 2.

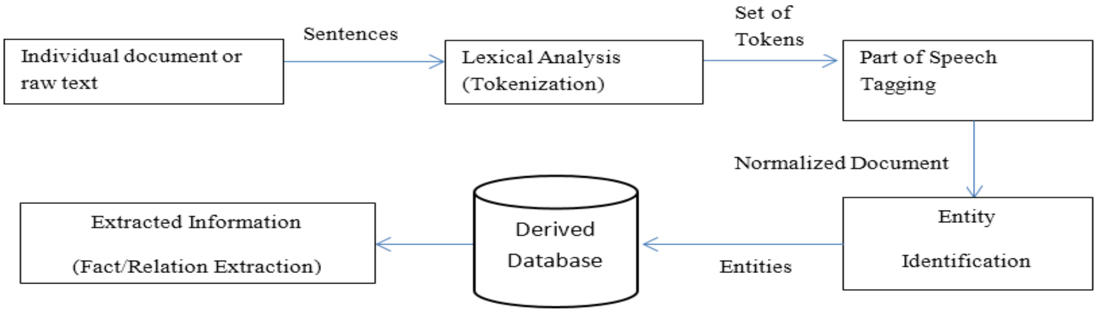
Individual document or raw text
Sentences
Lexical Analysis (Tokenization)
Set of Tokens
Part of Speech Tagging
Normalized Document
Entity Identification
Entities
Derived Database
Extracted Information (Fact/Relation Extraction)

Figure 2. Process of Text Mining

C) Parts-of-Speech Tagging

Parts-of-speech tagging is a piece of software that reads text in some language and assigns parts of speech to each word such as noun, verb, adjective to name a few. Generally computational applications utilize more fine-grained Parts of speech tagging include tags like 'noun-plural'. Dictionaries have category or categories of a particular word which implies that a word may belong to more than one category. For example, ‘Run’ is both a noun and verb. Taggers employ ‘Probabilistic Information’ to solve this ambiguity.

D) Named Entity Recognition

Named Entity Recognition is a subtask of information extraction that seeks to locate and classify Named Entities in text into pre-defined categories such as the names of persons, organizations, locations, expressions of times, quantities, monetary values, percentages, etc. For an instance, Robert bought 500 shares of Accenture Corporation in 2008. In this, a person name consisting of one token, a two-token company name and a temporal expression have been detected and classified. Hand-crafted grammar-based systems typically obtain better precision .In the current, statistical models are preferred as this approach initially uses training data against the model, followed by preparation of statistics [4]. These statistics are then used against real documents. Named Entity Recognition is also offered as a solution to NLP problems in various organizations like Stanford University [5]. Moreover it is also employed in libraries and Java platform to identify names and Entities. For Example, the newsfeed “enjoying U.S. weather at Texas with MonaLisa” will extract entities like weather, texas and Monalisa.

E) Named Entity Disambiguation

The task of linking the identity of entities available in the text is referred as Named entity disambiguation. However, it is distinctive from named entity extraction as it identifies not the occurrence of names but their reference. It needs a Knowledge Base of entities to which names can be linked.

F) Fact/Relation Extraction

Once named entities have been identified in a text, we can then extract the relations or facts that exist between specified types of named entity. The objective of the fact extraction is to detect and distinguish the semantic relations between entities in text or relations and fill it in a predefined template using the entities.

G) Word Sense Disambiguation

This is an open NLP and ontology subject that identifies the correct sense of the word in a sentence where multiple meanings of the word exist. It’s easy for a human to understand the significance of a word based on the basis of its background knowledge of the subject. However, identification the aspect of the word is difficult for a machine to understand. This methodology provides a mechanism to diminish the ambiguities of words in the text [7][8]. For example: Word Net is a free lexical database in English that contains a large collection of words and senses

H) Sentiment Analysis

Sentiment Analysis is an NLP process which identifies, extracts, enumerates the attitude of the user to the information that is provided by the user in a free form text. A text collection could show various sentiments which can be positive, negative or neutral. Sentiment Analysis is extensively used in processing survey form, online reviews and social media monitoring. It returns the identified sentiment with a numeric score from 1.0 to -1.0 where 1.0 means strongly positive and -1.0 means strongly negative [9]. For Example, “I love it” with score 0.8 means a strongly positive analysis for the newsfeed or blog. A practical application of this can be in a typical e-commerce website. Famous or ‘Top Rated” products are likely to attract thousands of reviews and this may make it challenging for prospective buyers to track relevant reviews that may assist in making decision. Sellers use sentiment analysis for there to decide relevant review and ignore the misleading reviews present to reviewers. A 5-star scale rating with five signifying best rated while one signifies poor rating.

2.1 OPEN SOURCE NLP LIBRARIES

NLP libraries are the algorithmic edifice of NLP in real-world applications. It provides a free API to setup or provision servers and infrastructure.

- Apache OpenNLP: It is an open source machine learning toolkit that provides natural language text. It provides services like tokenizers, summarization, searching, part-of-speech tagging, named entity extraction, translation, information grouping, natural language generation, feedback analysis and more. It provides a command line interface with some predefined models where models are trained and evaluated.
- Natural Language Toolkit (NLTK): It is a leading Python library that provides modules for processing text, classifying, tokenizing, stemming, sematic reasoning, parsing, and more. It provides user friendly interfaces over 50 corpa and lexical resources such as WordNet.
- Standford NLP: It is a suite of NLP tools that provide part-of-speech tagging, the named entity recognizer, coreference resolution system, sentiment analysis, and more. It provides statistical NLP, deep learning NLP, and rule based NLP tools which are broadly used in industry, academia and government.
- MALLET: It is a Java package that provides Latent Dirichlet Allocation, document classification, clustering, topic modeling, information extraction, and more.

3. CHALLENGES IN NLP

1. Informal language: Social Network users Poste texts in an informal language which is noisy include lack punctuation, misspellings, uses non-standard abbreviations, capitalization, and do not contain grammatically correct sentences.
2. Part-Of-Speech tags make the Information Extraction from social network more challenging.
3. Short contexts: Social Networks poses minimum length like Twitter. Thus, the user uses more abbreviations to precise more information in their posts. It is difficult to disambiguate mentioned entities due to the shortness of the posts and to resolve co-references among the feeds.
4. Noisy sparse contents: The users' post on social network does not always contain useful information. To purify the input posts stream, Filtering is required as a pre-processing step
5. Information about entities: People normally uses social Network to express information about their daily routine, happenings or about events and thus the entities are not contained in the knowledge Base. The Information Extraction approaches link the entities involved in the extracted information to a

Knowledge Base. There is a need of new Suit of Information Extraction from Social Network posts.

6. Uncertain contents: Not all information is trustworthy on the social network. Information contained in the users' contributions is in conflict with other sources and sometimes untrustworthy. The uncertainty involved in the extracted relations/facts is difficult to handle.

4. TEXT MINING

Text mining refers to the employment of data mining techniques which automatically discover and extract information from unstructured text documents and services. NLP is an attempt to extract meaningful information from free text. Information is gathered from large scale databases with the help of traditional data mining commonly known as warehouses. Then this data mining aids in extracting information. The aim of the paper is to discuss such approaches to form a super smart system which would analyze Social Network information.

Searching with the help of text mining is a way of retrieving and searching on a social searching engine that mainly searches user-generated content such as news, videos and images related search queries on social media like Facebook, Twitter etc. The text mining approach consists of four steps, which include Data Collection, Preprocessing, Generalization and Analysis

A) Data Collection

This is the process of gathering and measuring information in a systematic manner, which then enables one to answer relevant questions and evaluate outcomes. It deals with the challenge that updated information can be searched for a couple of days and the previous ones are not found. This happens as it becomes expensive if they store excess Data. There are huge numbers of users who access historical data at a particular time and it becomes difficult and expensive for social network to gather large amount of data. So, Summarization maintains all important data and further discards the insignificant data.

B) Pre-processing

This step refers to the processing of raw data to deliver a podium for data analysis. The significant purpose of this step is to classify raw sentences into sentences which can be read by the machine. The text is cleaned and delimiters are removed with the help of some pre known list of stop words which are not useful to classify the meaning of the sentence. The text and its characteristics are pointed in an attribute value table. Users enter the social text in a free form and therefore it is a challenging task to classify that data. Just to be sorted out from this challenge, part-of -speech tagging and Named Entity Recognition are used [9]. But this approach is costlier than others in terms of needful storage. For an accurate interpretation, attributed value table is important.

C) Generalization

This step involves the multiple patterns at the text of the preprocessed texts. It deals with developing algorithms to ascertain stimulating, unforeseen and unusual information form the patterns in the text document. One of the common tasks that occur is referred to as Apriori [12]. Frequent behaviors of persons or entities are recognized in the dataset. It identifies the inherent regularities in the data. This method was initially introduced in order to analyze customer buying behaviors from retail transaction databases.

Association, correlation, classification, cluster analysis form the strong foundation of data mining chores. For example, finding a strong correlation between two users A and B, of the connection A ⇒B, indicates that user that likes a product were also likely to be preferred by his friend B, so using this rule company can make decision to sell product to B who hold strong friendship relation with A. Finding the user's opinion about a topic is another example. This can be done by using sentiment analysis to determine how the topic is discussed on Twitter or other social networking sites.

D) Analysis

It deals with the validation and interpretation of the generalized data pattern. Density, Centrality, indegree, outdegree, and sociogram are the major terminologies to analyze the social network [6]. Degree identifies the "connections" between the users. Centrality determines the behaviour of individual user in the associated network. Indegree and outdegree are the measures of centrality. Indegree claims the individual user as the central identity; centrality of other users is based on their relation to the user whereas in outdegree the interaction of the user with others is the main focus point. A sociogram is the representation with the limited boundaries of the connection in the network is the point of analysis.

4.1 Applications of Text Mining in Social Network

Some applications of text mining in social network are [3]:-

1) Keyword Search

A set of keywords are used to identify the social network nodes which are close to the query result. Content and Linkage behavior plays an important role in order to determine the query output. It provides an effective method for accessing structure data. Query Semantics, Ranking Strategy and Query Efficiency are the major concerns to perform keyword search in social networks.

2) Classification

The nodes in the social network are associated with labels which are used for classifying the network. There are numerous algorithms available for classification of text from the content. The major issue in classifying the labels of the social network is the non-standard vocabulary and the noisy information associated with the labels as the labels of the social network are often sparse.

3) Clustering

It is the area where set of nodes are used to determine the similar content for the evolution of clusters. There are various clustering algorithm have been proposed which uses deviations of multi-dimensional data clustering techniques. K-means is the widely adopted technique where initial value of k is specified and clusters are built iteratively around that value. The algorithm iterates over parameters until an effective solution is reached. Linkage of clusters is an important factor and when combined with content can classify the social network which results in better clusters. Heterogeneity of the social network is the major concern that makes the algorithmic design more difficult to provide a viable solution.

4) Linkage based Cross domain learning

The linkage information between multiple domains of social networks provides transfer of knowledge across various kinds of links. The major concern in this learning is the amount of training data available from multiple social networks. An effective learning process can be leveraged from the various domains of the social network.

5. CONCLUSION

This paper discussed NLP techniques for Social Network that can enhance the experience of the user in more interactive way. Traditional text mining techniques are not popularly used in social network monitoring. The combination of text mining and web mining techniques should be incorporated to analyze a social network monitoring system. NLP Techniques will enhance a user friendly search by the Social Network user while text mining encompasses the intelligence in the Social Network.

REFERENCES

[1] Bikel, D., & Zitouni, I. (2012). *Multilingual natural language processing applications: from theory to practice*. IBM Press pp 400.

[2] Xiang, R., Neville, J., & Rogati, M. (2010, April). Modeling relationship strength in online social networks. In *Proceedings of the 19th international conference on World wide web* (pp. 981-990). ACM.

[3] Aggarwal, C. C., & Zhai, C. (Eds.). (2012). Mining text data. Springer Science & Business Media.

[4] Bikel, D., & Zitouni, I. (2012). Multilingual natural language processing applications: from theory to practice. IBM Press pp 286.

[5] The Stanford Natural Language Processing Group. (2012, Nov 15). Stanford Named Entity recognizer (NER) [Online]. Available: http://nlp.stanford.edu/software/CRF-NER.shtml#About

[6] De Laat, M., Lally, V., Lipponen, L., & Simons, R. J. (2007). Investigating patterns of interaction in networked learning and computer-supported collaborative learning: A role for Social Network Analysis. International Journal of Computer-Supported Collaborative Learning, 2(1), 87-103.

[7] Bandyopadhyay, S. (Ed.). (2012). Emerging Applications of Natural Language Processing: Concepts and New Research: Concepts and New Research. IGI Global.

[8] Princeton University. (2012, Nov 10). WordNet: A Lexical database for English [Online]. Available: http://word net.princeton.edu/

[9] Louis, A. (2017). Natural Language Processing for Social Media.

[10] Liu, B. (2012). Sentiment analysis and opinion mining. Synthesis lectures on human language technologies, 5(1), 1-167.

[11] Kasemsap, K. (2016). Text Mining: Current Trends and Applications. Web Data Mining and the Development of Knowledge-Based Decision Support Systems, 338.

[12] Agrawal R, Srikant R (1994) Fast algorithms for mining association rules. In: Proceedings of the 20th VLDB conference, pp 487–499

[13] Lin, J., & Dyer, C. (2010). Data-intensive text processing with MapReduce. Synthesis Lectures on Human Language Technologies, 3(1), 1-177.

[14] M. Song and Y. Wu, Handbook of Research on Text and Web Mining Technology, Hershey, PA, USA: IGI Global, 2009, pp. 228

[15] Gupta, V., & Lehal, G. S. (2009). A survey of text mining techniques and applications. *Journal of emerging technologies in web intelligence, 1*(1), 60-76.