International Journal of Database Theory and Application
Vol.9, No.8 (2016), pp.39-48
http://dx.doi.org/10.14257/ijdta.2016.9.8.05

Automatic Social Media Data Extraction

Estelle Xin Ying Kee and Jer Lang Hong

School of Computing and IT, Taylor's University, Malaysia
estelle.kee@taylors.edu.my, jerlang.hong@taylors.edu.my

Abstract

Opinions, the key influencer of human behavior and activity is ranked as of one of the strong factors that determine the effectiveness of one's strategy and approach in terms of influential power and trend setting capabilities. This highlights the importance of sentiment analysis done upon the extracted data. Today, statistics have shown significantly that most opinions can be obtained via many social media platforms. Social media has provided a convenient platform for web users to comfortably share their thoughts and to boldly voice up. Having to process such huge amount of data, it is proposed that automated sentiment analysis is done when extracting social media data. Using an effective algorithm which produces meaningful information from raw data, the possibilities of venturing deeper into areas like decision making and influential thinking are simply limitless.

***Keywords**: Social Media, Data Extraction, Semantic Analysis*

1. Introduction

Snapshots of sentiment can be found abundantly online specifically in social media. With the massive research done on sentiment analysis (some referred it as opinion mining), it is said that the sentiment of a social media author greatly exerts influence on the sentiment of people that in within the author's social media network. In other words, with social media, the potential of one impacting the people within his/her network is a whole lot more when compared with face-to-face communication. This also takes into account some other factors such as the broadness of one's network and the frequency of the contacts' exposed to social media. Therefore, triggering the creation of tools solely to analyze the sentiments collected via social media (Figure 1) [14].

Driving researchers further within this area of study are the struggles found among the social marketers of their inability to link the data they gathered from social media to easy meaningful representation in forms of tuples and attributes. It is undeniable that today's trend of investing millions of dollars into social marketing raises the demand and expectations of the work done presented by social marketers. However, most social marketers do not maximize the potential of which social media data extraction can be done. According to statistics provided by Dave Lloyd, a senior manager of Global Search Marketing at Adobe Systems, fewer than 50 percent of digital marketers understand whether social is working for them. This indicates a lack of focus on understanding consumer sentiment and also the ignorance of seeking the most optimum tool used to extract social media data along with the application of sentiment analysis.

ISSN: 2005-4270 IJDTA

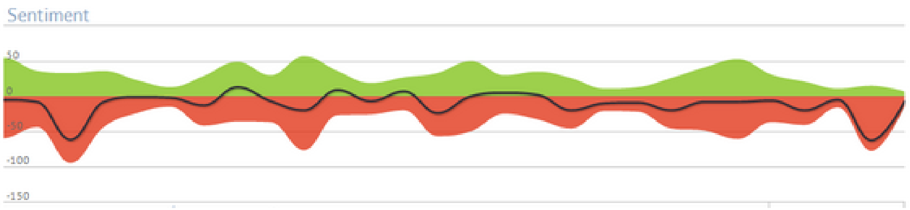


Figure 1. Sentiment Analysis by Coosto

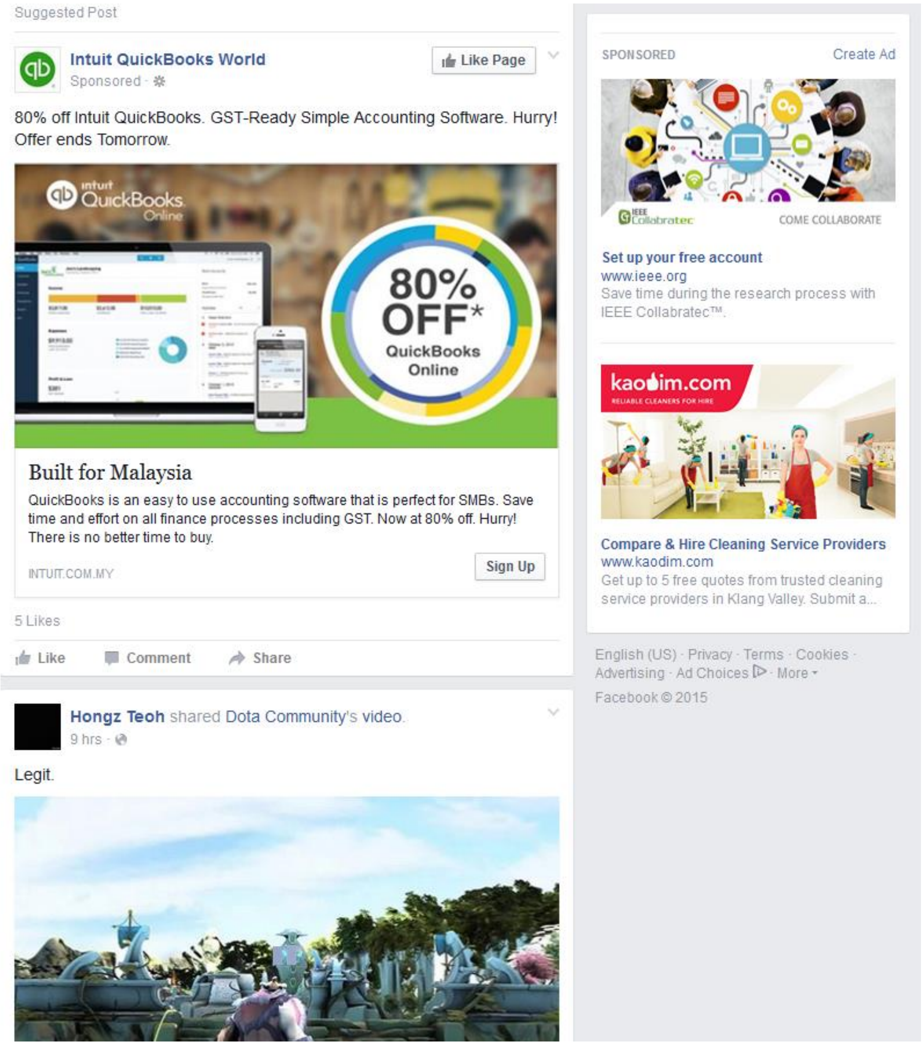
Therefore, the purpose of this paper is to enable social media data extraction to be done in a more efficient manner along with the sentiment analysis involved. The nature of sentiment analysis we have chosen for to suffice the objective of this paper is the unsupervised also known as the automated approach. With proven facts that the social media data today is strongly convinced to be growing exponentially and changing dynamically in a real time manner, the automated approach is the most realistic and efficient way to go about when extracting data from social media (Figure 2).

2. Related Work

Data that can be retrieved from the Web is often structured. These structured data on the Web are typically data records retrieved from underlying databases and displayed in Web pages following a noticeable pattern or template which can be seen in forms of repeatable segments. This has indeed motivated researches around the world to work on proposals in analyzing web documents and extracting the relevant information in structured formats [1]. These proposals are most commonly referred to as wrappers or as information extractors [4-12].

With studies done on the contemporary IE systems, the common categories of approaches used in data extraction are namely the manual approach, wrapper induction and automatic extraction [2]. Seeing the massive amount of social media data, automatic extraction is deemed the most ideal approach in handling data of such significant volume. Automatic extraction is described as an unsupervised approach that was introduced in 1998. This approach automatically seeks patterns or grammars from the given single or multiple pages before executing data extraction. One of the benefits of automatic extraction is the elimination of manual labeling effort and the ability of scaling up data extraction to a huge number of sites and pages [3].

In sentiment analysis, the objective is to discover all opinion quintuples in a given opinion document, D. Sentiment analysis contains a total of 6 main tasks [13]. With entity extraction and categorization being the first task, all entity expressions in D are extracted and synonymously categorized into entity clusters. Followed by the extraction of all aspects expressions of the entities and categorize these aspect expressions in their respective clusters. Continued on with opinion holder extraction and categorization as well as time extraction and standardization, the opinion is then determined whether it possess the traits of being positive, negative or neutral or it can be given a numeric sentiment rating. The sentiment analysis is then said to be completed with the final step of opinion quintuple generation.



Suggested Post

Intuit QuickBooks World
Sponsored ·

80% off Intuit QuickBooks. GST-Ready Simple Accounting Software. Hurry! Offer ends Tomorrow.

Built for Malaysia

QuickBooks is an easy to use accounting software that is perfect for SMBs. Save time and effort on all finance processes including GST. Now at 80% off. Hurry! There is no better time to buy.

INTUIT.COM.MY

5 Likes

Hongz Teoh shared Dota Community's video.
9 hrs ·

Legit.

SPONSORED

Set up your free account
www.ieee.org
Save time during the research process with IEEE Collabratec™.

Compare & Hire Cleaning Service Providers
www.kaodim.com
Get up to 5 free quotes from trusted cleaning service providers in Klang Valley. Submit a...

English (US) · Privacy · Terms · Cookies · Advertising · Ad Choices · More ▾
Facebook © 2015

Figure 2. An Example of Social Media Data

Opinion summarization on the other hand is comparatively subjective than factual information [13]. An opinion from a single source is often insufficient to make any conclusion. Therefore, in opinion mining, it is advisable to collect and analyze opinions gathered from a large pool of sources. Apart from including some of the key components of an opinion summary such as opinions about different entities and their respective aspects, it is important that a quantitative perspective is included as well. This being highlighted as 10% of the sources' positive opinion on a product or concept is very different from the remaining 90%.

3. Problem Formulation

The nature of social media data is said to be as one of the most diversified, massive yet dynamic form of data of which the we have ever encounter. With the continuous increase of frequency of users today expressing sentiments to the digital counterpart of the world – social media, it is found that many are interested to gain meaningful information out of it to enhance business processes, marketing strategies and so on. With various and ever increasing social media platforms out there, the presentation of semantics and information of which users provide are being represented in forms of variety. This inspires further research to be done in finding the appropriate algorithm to extract social media data to a fixed meaningful form even when the source is being presented in ways of different templates.

4. Motivation

Whenever the term “social media” is highlighted, it is automatically linked to people wanting to have a significant presence in the digital world, the urge of wanting to connect with people that matters and things of interest. Far beyond novelty, social media has been the prime focus of the ever-expanding group of netizens, desiring to master and utilize this platform to the fullest to achieve personal goals. Interested parties consists a majority of internet marketers which is determined to identify and understand clearly consumers’ feelings and behaviors towards brand and products. Not to forget, governing parties who are eager to win citizens’ hearts by understanding their expectations on future plans as well as feedbacks on current political policies and approach, to ease the customization process of a tailored-made election campaign. With the mentioned examples closely relating to sentiment analysis (also known as opinion mining), researchers realized the potential in which social media holds, in being a data source to many meaningful and valuable data which can be interpreted and presented in meaningful representations. Bringing us closer to the topic of data extraction, it is possible for us to extract the social media data using automated sentiment analysis. Being the most ideal choice to extract large amounts of data, using automated sentiment analysis also has its plus side of having a runtime of high speed whilst close to real-time. Besides, utilizing automated sentiment analysis as the data extraction tool incurred a lower cost compared to manual sentiment coded by human analysts, therefore benefiting the extraction process in a long run. With data extraction done using the automated analysis, we are confident that interested stakeholders such as government organizations and internet marketers would be able to benefit from this. Hence with this motivation, we strive to illustrate further in detail on how data extraction is being carried out using automated sentiment analysis.

5. Proposed Solution

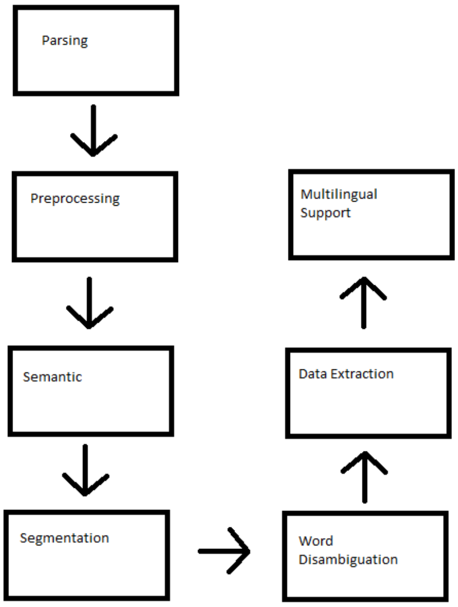


Figure 3. Flow Chart of Our Algorithm

Overview

There are various methods used to extract data from social media. Earlier methods utilize DOM Tree where the tree structure of the HTML page is analyzed and processed. Recent methods use visual cue where the visual boundary and size are used to measure the different regions, hence extraction is carried out based on the measured size. Lately, researchers use ontology tools to extract data where semantic properties of data is analyzed. Since majority of social media data is based on user's opinion, and they are unstructured in nature, we feel that sentiment analysis methods are able to extract this kind of data. To achieve this, we use WordNet semantic tool to check the semantic similarity of the social media content. The subsequent steps describe in detail our approach used for social media data extraction.

Webpage Parsing

We first parse the HTML page using a DOM parser and construct the DOM Tree accordingly. Once the DOM Tree is constructed, we parse through the DOM Tree in depth first manner, and for every block tag (*i.e.* div tag) detected, we measure the boundary of each block tag. Block tag with reasonable size will be considered potential regions and taken for consideration in the next step.

Preprocessing Stage

Punctuation are expressive elements used to create sense, clarity and stress in sentences. It helps user to understand better the different contexts of the sentence of which the author which to portray. However, in crafting the promotion model, theses accessories are not required, thus punctuation splitting which inclusive of white spacing is applied.

The collection of punctuation is fixed and known to the public. Hence, the project team decided to leverage on the regular expression settings to eliminate the punctuation, returning an array of String that are punctuation-free.

Spell check is to ensure the words are spelled correctly based on the defined-language which in the project context is the English language. With a high degree of ambiguity present in the human language alongside the variety of expressions and acronyms used, the spell checking procedure is indeed an important step to validate the input text provided by the merchants themselves. The spell check function will then flags words that are incorrectly spelled and replace the incorrect word via an autocorrect function with first preference of suggested word being applied. In achieving high accuracy, the spell check function requires a basis of reference to dictionary sources and semantic tools. With the spell check function being applied as part of the pre-processing, the input text is made easy for further processing.

In sentences, usually the words are of different forms that deviates from their root form. Stemming is a process to reduce inflectional forms and derivationally related forms of a word to its common base form. For example, “car”, “cars”, “car’s” will be stemmed to just “car”. Often referring to a crude heuristic process that removes derivational affixes, stemming is widely used in many application as it greatly reduces the complexity and variations of words that is to be further processed and thus increasing the effective of data analytics and machine learning algorithm that are in placed in later stages of the program.

Vector spaced scoring is a concept adopted to determine the size of a token by checking the highest frequency of appearance of the phrase or word and hence conclude the token size. This notion of technique greatly strengthen the reasoning of token sizing selected. For example, the term “German Shepherd”, it can be tokenized in 3 different ways, which are either “German”, “Shepherd” or “German Shepherd”. With ranking of the different arrangements of token sizing, we can then conclude that the highest frequency is the most dominant and more relevant to be selected as the token. This techniques is applied in many different contexts as it is proven to be effectively useful. With tokenization done, tokens are then parsed to the stage of semantic analysis.

DISCO Semantic Tools

For each potential region, we traverse through its tree structure, and for every text node detected, we perform semantic check on its content. We are under the assumption that social media content share similar semantic properties. Other irrelevant data such as menus, advertisements, and banners are not semantically similar. To check on the semantic properties of social media content, we use DISCO which provides a huge semantic tools and functions. We did a throughout survey on the available methods for semantic check and found that the algorithm of Jiang and Conrath is the best for our evaluation. Before we perform our semantic check, we perform POS tagging on the individual words, and we also stemmed the word to its base word. After that, we perform word similarity check and then we considered for Word Disambiguation using Adapted Lesk algorithm.

Webpage Segmentation

To segment a webpage efficiently, we use DOM Tree and its underlying tree properties. We first parse through the webpage and construct DOM Tree accordingly. Once we have obtained the DOM Tree, we then traverse the DOM Tree using depth first search. Once we have reached the leaf nodes (also known as text nodes), we then traverse upwards until we find a block tag (*e.g.* <div>, <table> tags). Then, we try to search for the remaining text nodes located in this sub tree and obtain the visual information of this tag. If the block tag contains sufficient information to be identified as a region (*e.g.* big enough in size to constitute a region, with sufficiently large contents such as text and images), we will then take this tag as a segment. Otherwise, we traverse upwards further until a suitable block tag is found. In the case where a root node is reached, we take the immediate block tag as the segment, otherwise that particular text node is not counted as belonging to a segment, hence discarded from our lists.

Semantic Relatedness

Once we have identified segments, we then process the segments and check for their semantic properties. We traverse through the DOM tree of this segment and then we tokenize the content of the text nodes into individual words. Once we have obtained a bag of words, we then use WordNet word similarity check to check whether two words are similar or not. If they are 75% similar, they are considered as semantically matched. We then stored keywords which contain similar semantic properties. Some text contains word disambiguation. For example, the word “interest” in the sentences “Interest in book” and “High interest rate in bank” are having entirely different meaning. For such a case, we use Adapted Lesk algorithm to differentiate the meaning between these two keywords. Adapted Lesk algorithm detects the semantic of two similar keywords by checking their neighboring words and matched those neighboring keywords with WordNet similarity check. Since the two sentences mentioned previously have highly dissimilar keywords (*e.g.* book, bank), it is concluded that the two sentences are not semantically similar. The procedure of matching keywords is then repeated for the remaining of the text contents.

Data Extraction

If a segment has a large number of text containing similar semantic properties, we can then safely identify that segment as the relevant region, hence extract it out accordingly. Our intuition is twofold. First, we notice that a user’s comments and posts are usually related to each other, the topic under discussion has always been similar, that is they are always related to the topic. Secondly, a region has big certainty to be identified as relevant if they have many similar keywords and contents. Whether these regions are menus, or advertisements, company banners, or even search results, developers usually enwrapped their codes to place similar information within a particular region. In the case where we have not identified a relevant region for extraction, we further partitioned these regions into smaller regions and similar procedure mentioned previously is carried out on these smaller regions to identify the relevant region to be extracted out.

Multilingual Support

The method we used previously does not cater for other languages other than English. To date, there exist numerous webpages written in different languages. The earlier version of WordNet caters only for English language. Recently, research is carried out where support for other languages has been incorporated into WordNet. This is a significant advantage to our work as multilingual support provided by WordNet can be used to analyze the semantic properties of text data written in various languages. To check for semantic similarity between keywords written in other languages, we need to implement the similarity methods in WordNet to cater for other languages. Fortunately, it is not difficult to map the implementation of Word Matching in English to that of other languages of WordNet as the functionalities provided by WordNet across other languages are almost similar though the accuracy returned by all these different methods may not be exactly similar. For example, a match between Cat and Dog in English WordNet may return 75% similar while that of Chinese WordNet may return 73% similar. Once we have implemented all the similarity check methods for WordNet written in other languages, we repeat the similarity check procedure used previously.

Once done checking similarity and ambiguity of the tokens, classification can then be done to clustered related tokens together. Classification enables groups of highly related tokens to be parked under one class. This greatly reduces complexity and increases the effectiveness of processing in later stages especially for the data cleaning module.

6. Experimental Tests

We conduct our experimental tests on a wide range of datasets. We collect a random sample of 200 pages from the deep web repositories for single language and another 200 pages for webpages written in different languages. We measure the effectiveness of our algorithm using precision and recall which are formulated as follows:

Recall=Correct/Actual*100

Precision=Correct/Extracted*100

Correct depicts the number of pages where the relevant region is correctly identified. Actual is the actual number of pages containing relevant region. Extracted depicts the number of pages where relevant region is extracted. We benchmark our work against the work of OntoExtract,[22] a state of the art tool which utilize multiple ontologies.

As shown in Table 1, our tool is highly accurate when extracting relevant region from the deep web. This is because our tool is able to identify relevant region from the webpage regardless of their layout, structure, and format. Unlike OntoExtract, we use WordNet similarity check and multiple stages check for extracting data, hence resulting in higher accuracy. Our method also works well for webpages written in different languages, as shown in Table 2. The accuracy of OntoExtract drops significantly, due to its inability to check for word similarity in webpages written in different languages. Only the exact word matching in OntoExtract is able to detect the similarity of keywords, which is a significant disadvantage as deep web contains many words which may not be exactly similar.

Table 1. Experimental Tests (Single Language)

Terms	Our system	OntoExtract [15]
Actual	200	200
Extracted	188	186
Correct	178	143
Recall	89.00%	71.50%
Precision	94.68%	76.88%

Table 2. Experimental Tests (Multiple Languages)

Terms	Our system	OntoExtract [15]
Actual	200	200
Extracted	186	134
Correct	174	104
Recall	87.00%	52.00%
Precision	93.55%	77.61%

We also measure the speed of our tool compared to OntoExtract. Our tool is able to run in just 150 ms for data extraction, compared to OntoExtract which runs at 2s for data extraction. This shows that our tool is useful for large scale data processing, such as comparative shopping lists, and meta search engine applications.

7. Conclusions

It is observed that the semantic element of social media data is having great influence upon the community especially in consumer behavior and influential thinking. With social intelligence being the prime focus of most organizations today, integration of social data and enterprise data are leveraged upon in producing their unique competitive edge respectively. With this publication in sight, we strongly believe that the automated semantic analysis proposed will be a great addition in building holistic yet strategic systems with reference to industrial standards and future needs.

References

[1] A. S. Hassan and C. Rafael, “A Survey on Region Extractors from Web Documents”, IEEE Transactions on Knowledge and Data Engineering, vol. 25, no. 9, **(2013)**.

[2] C. H. Chang, K. Mohammed, R. G. Moheb and F. S. Khaled, “A Survey of Web Information Extraction Systems”, IEEE Transactions on Knowledge and Data Engineering, vol. 18, no. 10, **(1999)**.

[3] B. Liu, “Structured Data Extraction: Wrapper Generation”, Web Data Mining: Data-Centric Systems and Applications, **(2011)**, pp. 363-423.

[4] J. L. Arjona, R. Corchuelo, D. Ruiz and M. Toro, “From Wrapping to Knowledge”, IEEE Trans. Knowledge Data Eng., vol. 19, no. 2, **(2007)**, pp. 310-323.

[5] C.-H. Chang, M. Kayed, M. R. Girgis and K. F. Shaalan, “A Survey of Web Information Extraction Systems”, IEEE Trans. Knowledge Data Eng., vol. 18, no. 10, **(2006)**, pp. 1411-1428.

[6] S. Kuhlins and R. Tredwell, “Toolkits of Generating Wrappers”, Proc. Revised Papers Int’l Conf. NetObjectDays Objects, Components, Architectures, Services and Applications Networked World, **(2002)**, pp. 184-198.

[7] N. Kushmerick and B. Thomas, “Adaptive Information Extraction: Core Technologies for Information Agents”, Agent Link, **(2003)**, pp. 79-103.

[8] A. H. F. Laender, B. A. R. Neto, A. S. da Silva and J. S. Teixeira, “A Brief Survey of Web Data Extraction Tools”, SIGMOD Record, vol. 31, no.2, **(2002)**, pp. 84-93.

[9] W. Meng and C. T. Yu, “Advanced Metasearch Engine Technology”, Morgan & Claypool Publishers, **(2010)**.

[10] I. Muslea, “Extraction Patterns for Information Extraction Tasks: A Survey”, Proceeding AAAI Workshop Machine Learning for Information Extraction, **(1999)**, pp. 1-6.

[11] S. Sarawagi, "Information Extraction", Foundations and Trends in Databases, vol. 1, no.3, **(2001)**, pp. 261-377.
[12] J. Turmo, A. Ageno and N. Catala, "Adaptive Information Extraction", ACM Computing Surveys, vol. 38, no. 2, **(2006)**.
[13] B. Liu, "Sentiment Analysis and Opinion Mining", Morgan & Claypool Publishers, **(2012)**.
[14] L. Kroonenberg, "The Importance of Sentiment Analysis on Social Media", Coosto, **(2014)**.
[15] J. L. Hong, "OntoExtract – Automated Extraction of Records using Multiple Ontologies", IEEE International Conference on Fuzzy Systems, and Knowledge Discovery, **(2013)**.