International Journal of Database Theory and Application
Vol.8, No.5 (2015), pp.233-240
http://dx.doi.org/10.14257/ijdta.2015.8.5.20

An Ontology Based Text Analytics on Social Media

Pankajdeep Kaur, Pallavi Sharma and Nikhil Vohra

GNDU, Regional Campus, GNDU, Regional Campus, GNDU, Regional Campus,
Jalandhar, Jalandhar, Jalandhar
pankajdeepkaur@gmail.com, pallavitalks18@gmail.com,
nikhil.vohra1212@gmail.com

Abstract

The amount of digital information that is created and used is progressively rising along with the growth of sophisticated hardware and software. In addition, real-world data come in a diversity of forms and can be tremendously bulky. This has augmented the need for powerful algorithms that can deduce and dig out appealing facts and useful information from these data. Text Mining (TM), which is a very complex process; has been successfully used for this purpose. Text mining alternately referred to as text data mining, more or less equivalent to text analytics, can be defined as the process of extracting high-quality information from text. Text mining involves the process of structuring the input data, deriving patterns within the structured data and lastly interpretation and revelation of the output. This paper provides outline on text analytics and social media analytics. At the end, this paper presents our proposed work based on ontology framework to cope up with excessive social media textual data.

Keywords: *Text Mining; Data Management; Data Analytics; Social Media Analytics; Text Clustering; Ontology Framework*

1. Introduction

Text mining, sometimes referred to as text data mining, nearly equivalent to text analytics, refers to the process of extracting high-quality information from text. Text mining in general involves the process of structuring the input text (typically parsing, adding up some derived linguistic features and the removal of others, and successive insertion/deletions into the databases), deriving patterns within the structured data and to end with interpretation and consideration of the output [1]. The word 'high -quality' in text mining generally refers to some combination of originality, significance and interestingness. Text mining is vast area in comparison to information retrieval. Text mining processes in general include document classification, clustering the document, building ontology, sentiment analysis, summarization, Information extraction *etc.* On the other side, information retrieval deals with crawling, parsing and indexing and retrieving documents [1]. With an iterative approach, an organization can use text analytics to expand insight into the content-specific values such as outlook, strength and significance of the data. Text analytics can be defined as the process of deriving information from text sources.

The quantity of textual data is ever growing. Conventional data mining techniques are not sufficient to evaluate unstructured data. So there is need to use some linguistic approaches. Text mining processes the unstructured data derives meaningful information. And thus, makes the information accessible to the various data-mining algorithms, including statistical and machine learning [2].

Big data can be defined as large volume of data that can be characterized by features such as velocity, volume and variety of information assets. Big Data is useless in a vacuum. Its prospective value is wide open only when it is used to force decision making.

ISSN: 2005-4270 IJDTA
Copyright ⓒ 2015 SERSC

To enable such verification based decision making, organizations need proficient processes to turn high volumes of diverse data into significant insights. The overall process of extracting insights from Big Data can be broken down into five stages [2], shown in Figure 1. These five stages structure the two main sub-processes such as: data management and analytics. Data management involves processes and supporting technologies to attain and store data and recover it for analysis. Analytics, on the other hand, refers to techniques that can be used to analyze and intellect Big Data. Accordingly, Big Data analytics process can be considered as a sub-process in the on the whole process of 'insight extraction' from Big Data. Figure 1 shows the 'insight extraction' process from Big Data.

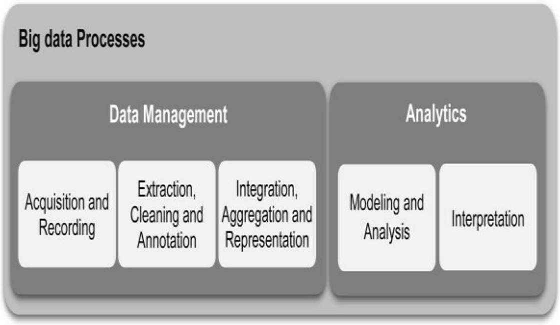
Big data Processes
Data Management
Analytics
Acquisition and Recording
Extraction, Cleaning and Annotation
Integration, Aggregation and Representation
Modeling and Analysis
Interpretation

Figure 1. Insight Extraction of Big Data Process

The various researches have focused in depth the different and important areas of text analytics and mining like social networks, industries, students' online interaction process business intelligence, digital libraries, security *etc.* In the following section we will study about social media analytics.

1.1. Social Media Analytics

Social media is defined as mobile-based and web-based Internet applications that permit the formation, access and exchange of user generated information that is far and wide easily accessible [3]. In social media analytics in addition to social networking media such as: Twitter and Facebook, the term 'social media' is also used to include really simple syndication (RSS) feeds, wikis, blogs and news, all yielding unstructured text that can be easily accessible through the web. Social media is primarily central for research into computational social science that examines and investigates questions using quantitative techniques such as, machine learning, computational statistics, data mining and simulation modeling [3-4]. This has led to the growth of numerous data services, tools and analytics platforms. Nevertheless, the ease of accessibility of social media data for intellectual research may be revolutionized appreciably due to commercial pressures.

1.1.1. Methodology and Requirements for Social-Media Analytics

The two major reasons of using social media for academic research are firstly, ease of availability and access to widespread data sets and secondly, access to tools that allow 'deep' analysis of datasets without the need to be able to program in a language such as Java [4]. It is important that researchers to have access to open-source social media data sets and amenities for experimentation. If not, social media research could become the restricted field of major companies and government agencies and including set of academic researchers that presides over private data from which they generate papers that cannot be replicated.

Copyright ⓒ 2015 SERSC

❖ Methodology

Research requirements can be grouped in 3 ways: data, analytics and facilities.

❖ Data

In order to conduct world leading researches, researchers need to access online historic and real-time data:

• **Social Network Media**—access to widespread significant data sets and furthermore real-time access to sources, perhaps with a minimum time delay (15 min) [4].

• **News Data**—access to momentous data and real-time news data sets, probably through the notion of 'educational data licenses' [4].

• **Public Data**—access to tatter and archived public data that can be obtainable through RSS feeds blogs and open government databases [4].

• **Programmable Interfaces**—researchers also require access to simple application programming interfaces (APIs) in order to scratch and store other accessible data sources that may not be collected automatically [4].

❖ Analytics

For the analysis of social media data we require:

•**Analytics Dashboards**—non-programming interfaces are required for giving 'deep' access to 'raw' data.

• **Holistic Data Analysis**—tools are required for collaborating and conducting analytics over numerous social media and other available data sets.

• **Data Visualization**—researchers have need of visualization tools where information that has been figure out can be visualized in some graphic form with the goal of communicating information evidently and effectively through graphical means. Figure 2 depicts the overall process of data analytics process. The process starts with analytics dashboard followed by data analysis and finally completing with data visualization.

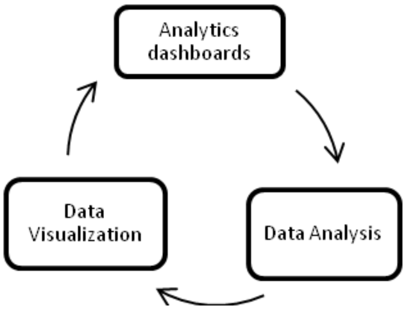

Figure 2. Data Analytics Process

❖ Facilities

Last of all, the complete volume of social media data being generated argues for national as well as international facilities that could support social media research work [4].

• **Data storage**—the amount of social media data, generated in day to day life is boundless, therefore needs to be addressed at national level. Storage is requirement for both primary data sources such as Twitter, but also for other sources collected by individual projects or to be used by future researchers.

•**Computational facility**—computational facilities [4] are required for:

a) Protecting access to the stored data;
b) Hosting the analytics and visualization tools;
c) Computational resources such as grids and GPUs are required for processing the data at the facility rather than transmitting it across a network.

Copyright ⓒ 2015 SERSC

1.1.1. Social Media Analytics Tools and Techniques

One of the key characteristics of enormous amount of user generated data and news content online is its textual disorder and high diversity. In order to analyze such data-sets natural language processing, text analytics and computational linguistics techniques are used to identify and dig out subjective information from source text [5]. The overall aim is to find out the outlook of a writer (or speaker) with respect to some subject matter or the overall contextual divergence of a document. In this section we will have over-view on the existing tool and techniques for analyzing social media data followed by the present tools and techniques that could replace existing ones.

1.1.2. Computational Science Techniques

Computational science techniques for the analysis of social media data includes machine learning, bag-of-words model semantic orientation *etc.*, [5]. Figure 3 shows the overview of existing computational science techniques that can be used for analyzing social media data. These existing techniques includes: computational statistics, machine learning, sentiment analysis, supervised learning methods *etc.*, [6, 7].

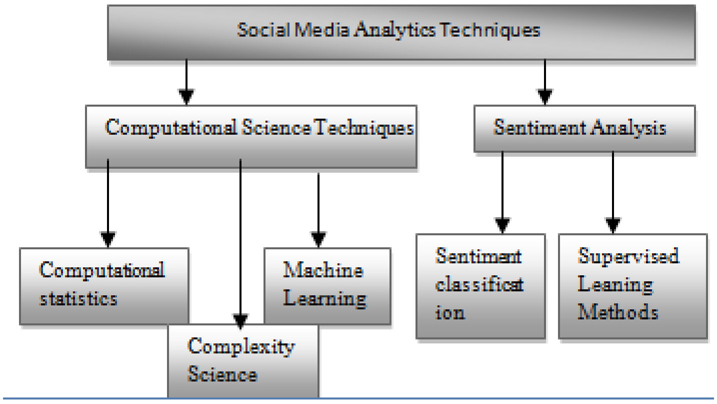
Social Media Analytics Techniques
Computational Science Techniques
Sentiment Analysis
Computational statistics
Machine Learning
Complexity Science
Sentiment classificat ion
Supervised Leaning Methods

Figure 3. Computational Science Techniques

2. Text Clustering

A clustering based approach to catch knowledge from text documents is referred to as data clustering. Document clustering or text clustering can be considered as a subset of the superior field of data clustering and text mining. The field of text clustering is based on the concepts of Information Retrieval (IR), Machine Learning (ML) as well as Natural Language Processing (NLP) field [8]. The overall process of document clustering can be defined as to automatically group or cluster a document into a list of meaningful categories, in such a manner that the documents in a category are similar or alike to each other and are dissimilar to documents in other categories [7]. It is one of the most essential tasks in text mining. There are number of techniques that are launched for clustering documents, so that straightforward document clustering to more demanding task such as production of granular taxonomies, document summarization, sentiment analysis for the scope of deriving high quality information from text.

Most of the information available in the internet in addition to intranets is in the form of text documents. In general, it is a requirement that these information sources are structured and sorted in some definite order so as to query for and retrieval of knowledge from these sources could be straightforward. These efforts can be viewed in two directions: First, in areas like information retrieval and text mining, researches and practitioners should look for categories of textual resources by means of some automatic methods [7-8]. These approaches can either *1)* predefine some metric on a document space to cluster the ‘nearby’ documents into significant groups of documents known as ‘unsupervised categorization’ or ‘text clustering’; or 2) they adapt to some metric on a document space so that predefined sample of documents can be assigned to a list of target categories manually. As a result, new documents can be assigned to labels from the target list of categories known as ‘supervised categorization’ or ‘text classification’ [8]; Second, researchers that are working under ontology should predefine some of the conceptual structures and assign metadata for the documents in such a way that confirm to these defined conceptual structures [8]. To gain advantages of both approaches, we will discuss an integrated approach of ontology learning and text mining framework.

3. Ontology Framework: An Approach for Data Retrieval

Ontology is a prescribed naming and definition of the properties, types and interrelationships between entities that exist for a particular domain [9].

There are several languages[9, 10] available for developing ontology like Common Algebraic Specification Language (CASL), Common Logic, Developing Ontology-Grounded Methods and Applications (DOGMA), Rule Interchange Format (RIF), Web Ontology Language (OWL). In this paper we used Web Ontology language (OWL).

3.1. Web Ontology Language

The OWL (Web Ontology Language) [11-13] was designed for those applications which need to process the content of information and not to just presenting information to the users. Web Ontology Language is mainly divided into 3 types:-

- OWL-Lite

OWL-Lite is the easiest version of the OWL family. It is easy to write and learn but it has one main disadvantage that is it restricts the expressiveness very much [12-13].

- OWL-DL

OWL DL is for those users who want the maximum expressiveness while performing complete computations. It also supports reasoning. It imposes certain restrictions. For instance a class can be a subclass of many classes but a class cannot be an instance of other class [12-13].

- OWL-FULL

OWL Full has full primitives of the OWL Language. Because of its large functionalities, it’s become difficult to use [12-13].

We used OWL-DL language to develop ontology for our system.

3.2. Proposed Ontology Framework

Figure 4 shows the diagrammatic view of our system. There are two interfaces available in the system. One for the user and another for the programmer. Different modules of the framework are explained as below:

- Database: All the information and data are stored in the database. The stored information is used to answer the user’s query. Any Changes made by administrator are simultaneously updates in the database.
- Browser: Web Browser acts an interface between user and Programmer. User can use OWL syntax for querying and developer can use SPARQL [14].
- OWL Ontology: In this part, rules are defined and developed. These rules define different ontology classes and interrelationships between them.
- Management Services: Different rules define different classes and through classes we provide services. Different types of services are available like Billing, Data Usage, and Authentication *etc.*

Copyright ⓒ 2015 SERSC

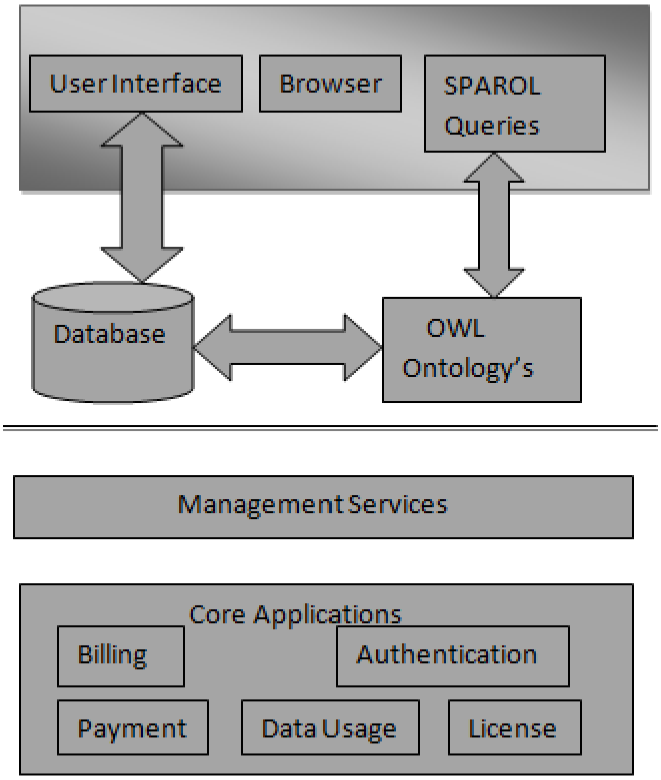

Figure 4. Ontology Framework

3.3. Ontology Classes

Some of the classes which we created in our ontology development are discussed as below:

- User: This class provides an interface to the users through which they can interact with the system *i.e*. they can submit their queries and get desired output. This class has two subclasses: RUsers and NRUsers. RUsers is class which keep track of the registered users and NRUsers class is for non-registered users
- Billing: Through this class, user performs their billing operations. This class has 3 subclasses: Credit, Debit, Netbank.
- Stock: This class contains record of the products available and non-available.
- Order: This class keeps the track of orders placed by the users.
- Offers: This class shows offers on the products available to the registered users and non-registered users.
- Categories: This class is intended to display categories of all products available in store. For example Mobiles and Tablets, Toys, Computers and Laptops *etc.*
- Supervisor: This class is for supervisor which maintains the database through this interface.

Figure5 shows hierarchical view of ontology classes of our system.

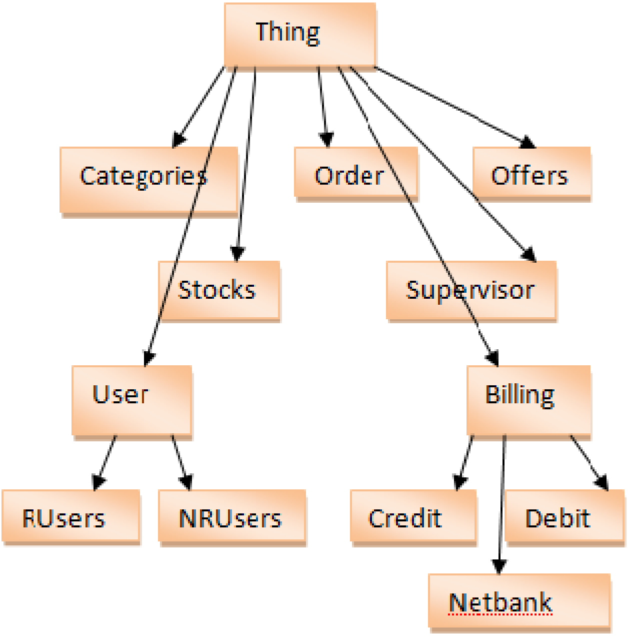

Figure 5. Hierarchical View of Ontology Classes

User use OWL syntax for querying something and Programmer uses SPARQL language for answering the query.

4. Case Study: E-Tourism

In this section, we discuss the applicability of ontology framework by illustrating how the different users interact with the system through the ontology classes.

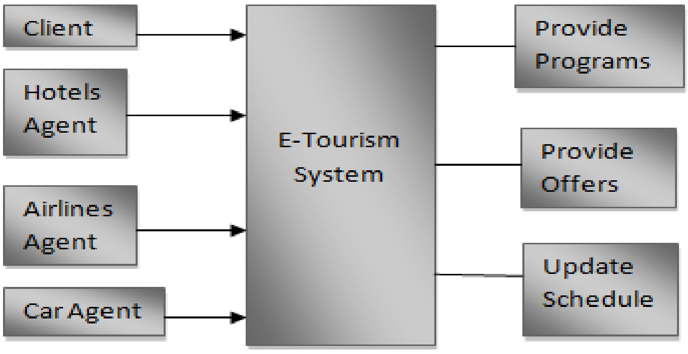

Figure 6. E-Tourism System with Required input and Output

Figure 6 shows the classes. Figure 6 shows the E-Tourism data inputs and output.

Figure 7 represents the data flow and control flow from Ontology classes to all agents. SPARQL has ability to detect and identify the agents based on type of request. The number of requests depends on the number of requests. In this case the SPARQL has a wide search function to search for the alternative agents with same feature and compare them to select best one.

Copyright ⓒ 2015 SERSC

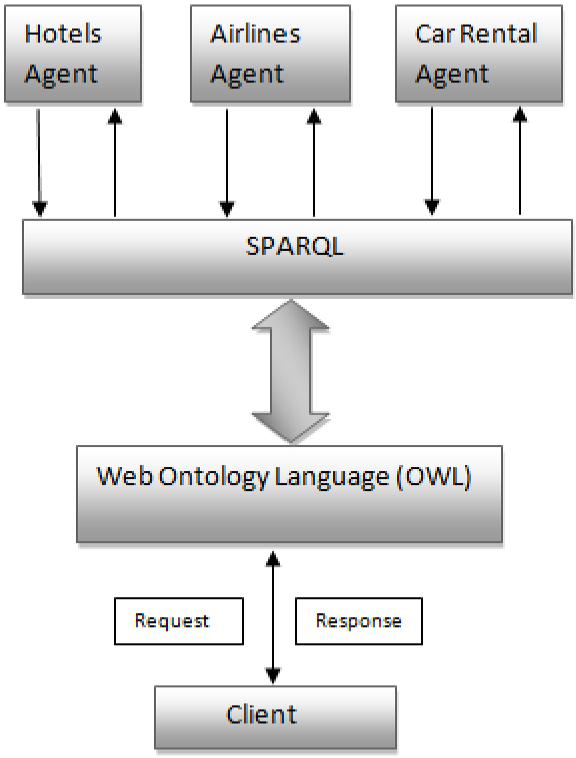

Figure 7. OWL and SPARQL Role on E-Tourism

5. Conclusion and Future Work

In this paper, we presented an ontology based framework that can respond to queries based on their semantics using OWL syntax. This framework provides users with good interface to interact with the system. Client uses OWL syntax and Programmer uses SPARQL language to handle the client requests. This system is scalable that is, if currently this system is capable of handling 1000 requests at a time and in future if we want to increase its throughput, then we can increase its throughput easily. Second, if we want to insert more functions in it in future, we can insert them easily .In future we will try further to extend this framework.

References

[1] Gartner, “Text Analytics”, **(2014)**.

[2] Nguyen, “Text mining and Network Analysis of Digital Libraries in Data Mining Applications with R”, **(2014)**.

[3] L. T. Berners, J. Hendler and O. Lassila, “The semantic web”, Sci. Am”, **(2001)**.

[4] He W., Zha S. and Li L., “Social media competitive analysis and text mining”, **(2013)**.

[5] Cioffi R. C., “Computational social science”, **(2010)**.

[6] Hirudkar A. M. and Sherekar S. S., “Comparative analysis of data mining tools and techniques for evaluating performance of database system”, **(2013)**.

[7] Murphy K. P., “Machine learning: a probabilistic perspective”, **(2012)**.

[8] Pang B. and Lee L., “Opinion mining and sentiment analysis”, **(2008)**.

[9] Han T. and Sim K. M., “An Ontology-enhanced Cloud Service Discovery System”, **(2010)**.

[10] Youseff L., Butrico M. and Silva D. D., “Towards United Ontology of Cloud Computing”, **(2008)**.

[11] Maniraj and Sivakumar, “Ontology Languages”, **(2010)**.

[12] Aref M. M. and Zhou Z, “The Ontology Web Language (OWL) For a Multi-Agent Understating system”, **(2005)**.

[13] Zhihong Z. and Mingtian Z, “Web Ontology Language OWL and its Description Logic Foundation”, **(2003)**.

[14] E. Prud’hommeaux and A. Seaborne, “SPARQL Query Language for RDF”, **(2006)**.

Copyright ⓒ 2015 SERSC