Proceedings of the Second International Conference on Computing Methodologies and Communication (ICCMC 2018)
IEEE Conference Record # 42656; IEEE Xplore ISBN:978-1-5386-3452-3

Data Mining with Linked Data: Past, Present, and Future

Rohit Beniwal, Vikas Gupta, Manish Rawat, and Rishabh Aggarwal
Department of Computer Science & Engineering
Delhi Technological University, DTU
Delhi - 110042, India
rohitbeniwal@yahoo.co.in, (vg29051997, manishrawatofficial, rishabhaggarwal1096)@gmail.com

Abstract- **Linked Data has emerged as a popular method for representing structured data. One of the prime aims is to convert today's web of documents into a web of data where the data is machine-readable as well as processable. This research paper focuses on the data mining techniques used for mining the raw data. However, these techniques are cumbersome and can be optimized using Linked Data. Hence, we discuss the data mining techniques with Linked Data that may play a pivotal role in future in extracting meaningful information from unstructured or semi-structured data.**

Keywords- Data Mining; Social Media Data Mining; Linked Data; Web of Data; KDD

I. INTRODUCTION

In every decade, the size of the storage devices shrinks, yet capacity increases dramatically. Due to the rapid advancement in data storage technologies, we are able to store an enormous amount of data [1]. Every day new and more data is uploaded over the web. However, this vast amount of data is useless without any fast and reliable method to process it. Data processing is a problem in itself. Obtaining useful information from large datasets is becoming a growing concern. As the volume of data increases, so is the ambiguity. Extracting relevant information would be easier if the data to be processed is available in a structured form. However, that does not happen in most of the cases. Mostly, one has to deal with unstructured or semi-structured data. For e.g. most of the data available on the web is in the form of HTML documents. HTML documents contain texts, audio, video, pictures, links, tables, etc. which are displayed by the web-browsers. Thus, making it difficult for computers to extract useful information from such plain text documents. Nowadays, for extracting meaningful information from unstructured or semi-structured data, data mining techniques are used very frequently. Therefore, this research article examines data mining techniques which are currently being practiced. However, traditional data mining techniques work well only with isolated data sets. Hence, we also discuss the data mining techniques with Linked Data that may play a pivotal role in future in extracting meaningful information from unstructured or semi-structured data.

The rest of the paper is organized as follows: Section II discusses about data mining and its needs followed by section III which talks about data mining with Linked Data; section IV elaborates about knowledge discovery followed by section V that converses about data mining through API. Additionally, section VI expounds about mining the web of Linked Data with tools; section VII discusses about semantic proximity with Linked Data. Section VIII concludes the research article.

II. DATA MINING & ITS NEEDS

"Data mining is the practice of automatically searching large storage of data to discover patterns and trends that go beyond simple analysis. Data mining uses sophisticated mathematical algorithms to segment the data and evaluate the probability of future events. Data mining is also known as Knowledge Discovery in Data (KDD)" [2].

For e.g., data mining aims to build a knowledge base which contains information that may even be useful to small businesses like local restaurants and stores for marketing purposes. The information posted on social networking sites can prove to be very useful to these local businesses. Users' comments sometimes include their views and opinions on places they visit and whether they had a good or bad experience. The problem arises because these comments and posts are available in plaintext and not in formal languages.

To resolve above problem, text mining techniques are required. They are used to extract useful information and keywords, and then can be converted to the formal knowledge database. Comments are analyzed for the name of the businesses, positive or negative feedback. Other information about users like age, gender, location, etc. is extracted which might give business owners additional input while coming up with marketing strategies. Further challenges may be filtering out fake reviews which are paid by business owners to make sure that they are positive. Information can be old or inconsistent.

Fortunately, there is a newer approach which is reliable, unambiguous and more efficient where we use Semantic Web to describe relationships and link isolated datasets.

"Semantic Web provides a common framework that allows data to be shared and reused across application, enterprise, and community boundaries" [3]. It is a machine processable "Web of Data" [4]. "The collection of Semantic Web technologies (RDF, OWL, SPARQL, etc.) provides an environment where the applications can query that data, draw inferences using vocabularies, etc." [5]. It is a development of World Wide Web in which data in a web page is structured and tagged in such a way that it can be read directly by computers [6].

The Semantic Web uses formal languages like Resource Description Framework (RDF) and Web Ontology Language (OWL) to define information semantically. RDF is a standard model for data interchange over the Web. The OWL is a Semantic Web language designed to represent rich and complex knowledge about things, groups of things, and relations between things [7].

Semantic Web languages make the open-world assumption. The absence of a particular statement within the web means, in principle, that the statement has not been made explicitly yet. In essence, from the absence of a statement alone, a deductive reasoner cannot (and must not) infer that the statement is false [8].

Linked Data lies at the heart of what Semantic Web is all about: large scale integration of, and reasoning on, data on the Web [5].

III. DATA MINING WITH LINKED DATA

Data mining not only observes the data but also relationship among the data. Data mining techniques include association, classification, clustering, prediction, sequential patterns, and decision trees.

Traditionally in data mining, analyst selects the relationships among the data, and this is a major disadvantage of classical data mining technique. The relationships that are provided by the analyst may be wrong or insufficient to determine overall behavior among the data. This leads to ambiguity and errors. In case of Linked Data, assumptions can be made without actually knowing them, in case of open world assumption. Thus, it enables data mining to have set of interconnections between different datasets. Interlinking is provided by semantics.

The techniques, which are currently used for data mining consider each dataset as an isolated source of data. However, this classical technique is fast but not reliable because it can result in ambiguity and may not emphasize additional meaningful information. This limitation can be overcome by use of the new field of Semantic Web known as Linked Data. As we know that "Semantic Web refers to an extension of the current Web that provides an easier way to find, share, reuse, and combine information. It is based on machine-readable information and builds on XML technology's capability to define customized tagging schemes and RDF's flexible approach to represent data" [9]. Moreover, Linked data is defined as a term used to describe a recommended best practice for exposing, sharing, and connecting pieces of data, information, and knowledge on the Semantic Web using URIs and RDF [10]. Here "Uniform Resource Identifier (URI) is a string of characters used to identify a resource" [11].

In order to use our knowledge about Linked Data in data mining, we must represent the data using semantic ontologies [12], 13]. This requires three steps to be followed. First, data is trained - training means data is initially observed. The interrelations between several datasets are looked into and are modified. During training, functional dependencies and associations among data are performed and applied. Different types of data are put into relevant classes. Second, the test is performed on our trained data. Several test cases are fed in, and outputs are checked with desired outputs. Third, check validation is done, i.e., requirement of the system is compared with the machine outcome, and if everything works fine, then our data is converted into Linked Data.

If related data is properly linked, knowledge about the data can be obtained. Knowledge Discovery and Data Mining (KDD) is multidisciplinary area that focuses on techniques for obtaining useful knowledge from data. The rapid pace at which the data has grown online has created an extensive need for KDD methodologies.

IV. KNOWLEDGE DISCOVERY

Knowledge plays a very important role in data mining. Knowledge can be extracted from the data or be fed into the system manually. It may already be in the database and can be extracted directly using certain techniques and algorithms. Sometimes data is not present in the first place, so we have to incorporate external data into our system. There may be the case when data is not present physically anywhere, but it is only feed in manually by the data analyst. Knowledge discovery is a post condition of data mining. The entity available after data mining is knowledge itself.

Semantic Web and Linked Data can be easily applied for the process of knowledge discovery. In order to get a Semantic Web of data, we apply Linked Data ideologies to the data mining process. This process includes six steps which are as follows.

A. Identification of Data and Datasets

The first step is the selection of relevant datasets and removal of redundant and useless ones. Data may be selected based on class in which it resides, operation or by any of its properties. Here data are classified into datasets and may be given some properties. Data is stored in relational databases.

B. Preprocessing Step

This process creates links between the related datasets. Two datasets are linked with some attributes which may be used later in the process of data mining. Relationships are currently limited to datasets. Web Ontology Language (OWL) will be used to develop ontologies [14], [15].

C. Modification of Data

Data is then transformed from relational form to graphical form having data as links and connections as

relationships. The graph can be directed or undirected; it depends on the nature of underlying relationships. Sometimes the number of datasets or nodes in a graph are limited. For example, when we want to show recommended settings to the user in social media, data from previous experience are gathered, and only some definite amount of relevant data is shown to the user. SPARQL queries can be performed to get relationships. SPARQL enables users to query information from databases or any data source that can be mapped to RDF[16].

D. *Mining*

Once the data sets are linked using semantic technologies, new information can then be deduced easily. As the data is mined, it is ready to be converted into knowledge.

E. *Gaining Knowledge*

Obtained data using data mining is now interpreted and evaluated to form our Knowledge. Moreover, this knowledge may be used further or be used directly by a user at runtime.

Querying of data can be done efficiently through the use of Application Programming Interfaces (API's) rather than SPARQL as using API's is simpler than SPARQL. Data integration is a critical requirement as data from isolated, scattered sources need to be integrated for making a unique knowledge base and information retrieval. Such a knowledge base would facilitate easy and effective data retrieval through data mining techniques. Although existing technologies can query data sets, the query requires explicit construction. APIs overcome this limitation by providing interfaces for intermediating query processing.

V. DATA MINING THROUGH APIs

By using API, data integration can be made better. APIs are a new solution for accessing the data sets for information retrieval. APIs allow encapsulation or data hiding because these data sets are accessed through a well-defined interface. Further, more developers are familiar with APIs than SPARQL or crawlers.

There is a basic solution for data integration called Linked Data Application Architecture (LDAA).

A. *Linked Data Application Architecture (LDAA)*

As shown in Fig. 1, different datasets or knowledge bases are converted into RDF data using Linked Data wrapper or Resource Description Framework in Attributes (RDFa). RDFa is a W3C recommendation that adds a set of attribute-level extensions to HTML, XHTML and various XML-based document types for embedding rich metadata within Web documents [17]. This converts the entire datasets into what is called the web of data. Data from this web of data is mined and integrated through different modules including web access module, vocabulary mapping module, identity resolution module, and quality evaluation module. This integrated data is finally queried using SPARQL.

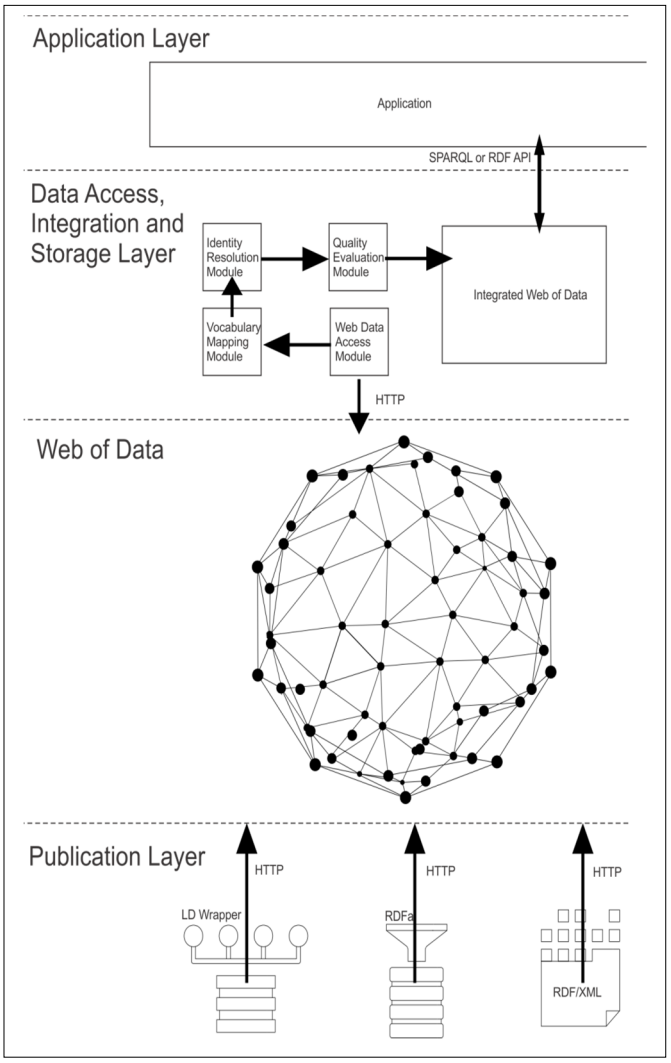

Fig. 1 - Linked Data Application Architecture

However, the LDAA concept cannot answer questions like which vocabulary should be chosen to represent the final integrated data.

B. *LDAA using APIs*

In this model, the API defines the type and nature of access methods that can be used by the user to mine linked data. The API hides the details of the access methods like its contents and how it passes the parameters. It does so by changing the data access, integration and storage layer by API mediator module, declarative API definition and API documentation.

This architecture is a significant improvement over classic LDAA as it successfully answers all the questions left open by LDAA.

C. *An Example - The Open Phacts Discovery Platform*

The open PHACTS platform, which is a Linked Data integration system for pharmacological data, is an extension of LDAA using APIs. The open PHACTS platform

combines 11 datasets and provides collective access to them.

Linked data also provides a way for those who are not specialized in computer programming or for those who know programming but still want things simpler.

VI. MINING THE WEB OF LINKED DATA WITH TOOLS

Although significant progress is made in the field of constructing and maintaining Linked Data, a versatile tool for deriving additional implicit knowledge by data mining is still missing. Rapid Miner linked open data extension is a tool that hooks into the Rapid Miner platform and allows for data mining without expert knowledge in SPARQL or RDF. It produces a set of operators for augmenting existing data sets with additional attributes from open data sources.

Many social networking websites with professional and non-professional orientations are serving as data sources for big data analytics. As the data available on social platforms is unstructured, more sophisticated methods need to be developed for social data mining.

VII. SOCIAL MEDIA ANALYTICS USING LINKED DATA

Since the data on social media platforms is in unstructured textual form, it is processed by data mining techniques like - folksonomy, concept-level sentiment analysis, etc. Folksonomy is a user-generated system of classifying and organizing online content into different categories by the use of metadata such as electronic tags [18].

Semantic technology provides background knowledge, context, interoperability, and is accepted as data expressivity standard.

A. Social Media Analytics Framework

For instance, Employers have started using LinkedIn for hiring people. Thus, an ocean of jobs related data is available.

A challenge here is to transform LinkedIn's data to knowledge base using semantic ontologies [19]. The framework proposed has the following components - first component is associated with data extraction, second component preprocesses unstructured textual data with basic text mining techniques. To analyze employment trends from LinkedIn ,we extract various relevant details like: job title, location, employer, etc. through data extraction algorithm. The derived data objects are converted to JSON-LD with relevant context.

B. Data Extraction, Transformation and Analysis

Extraction and mapping of data attributes are performed and using text mining techniques, information can be deduced.

The transformation phase converts resulting dataset into the semantic technology compatible format. Derived dataset is converted from text to JSON.

We prefer this semantic approach as it offers the open world assumption ,that is, facts that are not known are not said to be false. Moreover, new data facts added to the existing model can be considered in the analytical process.

VIII. CONCLUSION

In this research article, we discussed various techniques that are used for mining, processing, and querying the data along with establishing relationships among the data. We also examined Rapid Miner, which is one among the various tools that are available for mining the data. We also discovered that APIs can be used to make data mining easier and in some cases more efficient. Also, data mining through the use of APIs does not require extensive knowledge of SPARQL. Moreover, we also discussed some of the techniques used for social media data mining. At last, we conclude that data mining through the use of Linked Data has emerged as one of the most efficient ways of extracting and deducing useful information.

REFERENCES

[1] Data Storage -- then and now: https://www.computerworld.com/article/2473980/data-storage-solutions/data-storage-solutions-143723-storage-now-and-then.html#slide2 (Last accessed date: January, 2018)

[2] What is Data Mining?: https://docs.oracle.com/cd/B28359_01/datamine.111/b28129/process.htm#CHDFGCIJ (Last accessed date: January, 2018)

[3] W3C Semantic Web Activity: http://www.w3.org/2001/sw (Last accessed date: January, 2018)

[4] T Berners-Lee, J Hendler, O Lassila. "The semantic web." Scientific american 284.5, pp 34-43, 2001.

[5] Linked Data: https://www.w3.org/standards/semanticweb/data (Last accessed date: January, 2018)

[6] Semantic Web:http://www.dictionary.com/browse/semantic-web?s=t (Last accessed date: January, 2018)

[7] OWL: https://www.w3.org/OWL/ (Last accessed date: January, 2018)

[8] Open World Assumption: https://en.wikipedia.org/wiki/Open-world_assumption (Last accessed date: January, 2018)

[9] Semantic Web: https://www.webopedia.com/TERM/S/Semantic_Web.html (Last accessed date: January, 2018)

[10] Linked Data: http://linkeddata.org/ (Last accessed date: January, 2018)

[11] URI: https://en.wikipedia.org/wiki/Uniform_Resource_Identifier (Last accessed date: January, 2018)

[12] M.P.S. Bhatia, R. Beniwal and A. Kumar, "An ontology based framework for automatic detection and updation of requirement specifications." In Contemporary Computing and Informatics (IC3I), 2014 International Conference on, pp. 238-242. IEEE, 2014.

[13] M.P.S. Bhatia, R. Beniwal and A. Kumar. "Ontology based Framework for Ambiguity Detection software requirements specification." Computing for Sustainable Global Development (INDIACom), 2016 3rd International Conference on. IEEE, 2016.

[14] M.P.S. Bhatia, R. Beniwal and A. Kumar. "Ontology based framework for reverse engineering of conventional softwares." Computing for Sustainable Global Development (INDIACom), 2016 3rd International Conference on. IEEE, 2016.

[15] M.P.S. Bhatia, A. Kumar, and R. Beniwal, "Ontology Based Framework for Automatic Software's Documentation." In Computing for Sustainable Global Development, 2015 2nd International Conference on, pp. 725-728. IEEE. 2015.

[16] What is SPARQL: https://ontotext.com/knowledgehub/fundamentals/what-is-sparql/ (Last accessed date: January, 2018)

[17] RDFa: https://en.wikipedia.org/wiki/RDFa (Last accessed date: January, 2018)

[18] Folksonomy: http://www.dictionary.com/browse/folksonomy?s=t (Last accessed date: January, 2018

[19] MPS Bhatia, A Kumar, and R Beniwal. "Ontologies for software engineering: Past, present and future." Indian Journal of Science and Technology 9, no. 9, 2016.