A Survey on NLP Techniques to Extricate Semantics from Dataset

[1]Shweta S Aladakatti
Department of Computer Science & Engineering
Presidency University, Bengaluru, India

[2]Senthil Kumar S
Department of Computer Science & Engineering
Presidency University, Bengaluru, India

Abstract:- **Semantic Web helps to interconnect data sources on the Web, thereby presenting a global database of Web resources. Classification using similarity identification of data among different resource is one of the major challenge in semantic web. However, as the internet collects an increasing amount of unstructured data, the Semantic Web has the issue of keeping up with new and updated content. Natural language processing procedures and executions have been exhibited to be successful in defeating the difficulties of giving importance to huge volumes of unstructured information. There are few NLP techniques used to classify the data, such as Bag-of-words strategy, the TF-IDF method, the NER procedure, the LSA procedure, and the LDA procedure are completely shown in this work. The examination delivers either a comparable quality fulfillment or the ID of the importance on this unstructured information, contingent upon the method picked. NLP draws near, on the other hand, can possibly be valuable since they carry bits of knowledge into the information and make it more comprehensible. Accordingly, more easy to use human-machine interfaces and applications will be made.**

Keywords:- Semantic web, natural language processing, bag-of-words, TF-IDF, Named Entity Recognition, Latent Semantic analysis, Latent Dirichlet Allocation

I. INTRODUCTION

Essentially, the Semantic Web refers to a set of principles that govern how information is organized and shared on the internet. The semantic web and interlinking have been the focus of investigation since the beginning of the internet era [1]. As the number of apps, data, and users continues to rise at an alarming rate, controlling search results, demotion, and future data access becomes more difficult. The semantic web begins with the development of incipient and high-level access mechanisms based on semantics for the purpose of gaining access to online material and services [2]. Discovering the differences between data available in cyberspace is linked to navigation savvy, which has also been the subject of research in the communities of hypermedia systems and optimization, as well as in the communities of applications. The use of robust semantic layers on the web may be able to make human interaction with the web more interesting. Since 2008, there has been a surge in the number of research concentrating on the evolution of metadata at all scales, from the minute to the enormously massive. Various organizations, application-oriented businesses, and technically astute individuals have all contributed to enrich the semantic web with each data exchange that has occurred. One of the reasons for this level of semantic richness is the multiplicity of different types of internet businesses. It will not only help to improve the semantics of metadata by automating the process of semantic interlinking, but it will also aid in the acquisition of early and better representations of semantic enrichment. It is critical to educate oneself on how to interact with the unstructured data that is becoming increasingly available [2].

In the Well-organized information, the way links are statically dispersed can make a considerable difference in the quality of the information. Data mining, machine learning, data science, natural language processing, and other emerging technologies are proving to be extremely benign, as they speed up the process of extracting text from multiple datasets, acquiring existing relationships, and allocating the incipient data to the existing relationships, all of which are extremely beneficial to society. Using cross-functional technologies, it is feasible to achieve semantic enrichment as well as accurate interlinking of information. In some ways, semantics is akin to human language, and one of the most important goals of semantic success is to bring human and machine interaction closer together as a result of this success. The ability to process and generate human-readable interactions is inherited by NLP as a result of its inclusion in the artificial intelligence domain. Semantic analysis and interlinking are used in conjunction for semantic enrichment in order to discover the most relevant element of a text and to have a better understanding of the issues that are always being debated. Combining semantic analysis with natural language processing allows the computer to become exceedingly clever in its grasp of the information it is presented [5]. It is possible to strengthen the purpose of organizing unstructured data with a predicted use case and meaning to it even further by combining the two approaches. A variety of enterprises can benefit from this, among other things, by gathering client input, providing feedback or recommendations, and expanding their knowledge base. Natural language processing (NLP) is the automated understanding of information written in natural (human) languages rather than artificial languages (such as English, French, or Chinese) (such as programming languages). In some circles, it is referred to as Computational Linguistics (CL) or Natural Language Engineering (NLE). Natural language processing (NLP) applications range from simple text segmentation into sentences and words to more complex tasks like semantic annotation and opinion mining [15]. The Semantic Web is focused with adding semantics, or meaning, to Web data in order to aid computer digestion and modification of web pages. One of the most important aspects of the notion is the use of unique identifiers known as universal resource identifiers (URIs) to define resources (URIs) [2].

NLP techniques offer semantic enrichment of web data in a variety of ways, including automatically adding information about entities and relations and recognizing which real-world entities are being referenced in order to assign a unique URI to each cited entity. The scope of this research includes an investigation into the possibility for semantic analysis using natural language processing. In order to generate meaning from the unstructured data, the strategies are assessed using unstructured data acquired from the campus library. The goal of this work is to derive meaning from unstructured data. The study is organized in such a way that it provides a high-level overview of the need to improve semantic enrichment approaches in order to improve human-machine interaction and develop high-quality apps. This is performed by organizing the research in an orderly manner. The use of NLP methodologies resulted in the discovery of a semantically linked attribute, which is explored further in the section of the report devoted to result analysis [8].

II. LITERATURE REVIEW

Semantic interlinking works in view of human characterized ontologies [1], these ontologies are created and relegated by people, and people compose these ontologies in comprehensible dialects, consequently regular language handling will empower the human to convey and robotize interlinking. This permits a person to relate and figure out the characterized ideas, however in opposition to the assertion, machines can learn through ontologies and can have restricted expansion of connections [2]. NLP methods can help in better characterizing the semantic application, which is generally restricted by formal rationale closeness interlinking. Regular Language Processing strategies and procedures when conveyed inside semantic age approach can defeat the vagueness issues between assorted ontologies that are being utilized by semantic ward text, report or information extraction and characterization devices or projects [5]. A last and most critical boundary that is pertinent to the outcome of the devices or technique is that the people liable for the creation ought to know the business the instrument will be working for and not be exclusively determined by the experts [2]. Consequently the disclosure strategy should be created in a manner that would overcome any issues between catchphrases written in NLP and on the opposite end connected to the web descriptors that are given by NLP methods can assist with defeating the uncertainty issues between various ontologies that are being utilized by semantic Web administration depictions. Last, administration creation ought to be driven by individuals who realize business processes and not by professionals. Hence, end clients should have the option to find these Web administrations in view of catchphrases written in human language. In this manner, a revelation component should be created so that an extension between catch phrases written in a characteristic language [13].

2.1 Semantic analysis using Natural language processing

Natural Language Processing (NLP) is programmed handling of text written in regular (human) dialects (English, French, Chinese, and so on), rather than machine readable dialects such as programming dialects, to attempt to "get" it. It is otherwise called Computational Linguistics (CL) or Natural Language Engineering (NLE). NLP includes a wide scope of exercises, from low-level exercises, for example, fragmenting message into sentences and words, to intricate, undeniable level applications, for example, semantic explanation and opinion mining. The semantic web is tied in with adding semantics, that is, to the information on the web, so that site pages can be handled and controlled all the more effectively by machines. At the core of the thought is that assets are depicted utilizing remarkable identifiers, called Uniform Resource Identifiers (URIs). Assets can be elements, for example, "math subject", ideas, for example, "subject" or connections depicting how elements connect with one another, for example, "insights" [5]. NLP strategies give a method for improving web information with semantics, for example, naturally adding data about elements and connections and understanding which true elements are referred to so a URI can be doled out to every element. Before the dataset is based to the methods, the txt files are preprocessed and stored into data frames. The methods used in this paper are as follows:

2.1.1 Bag-of-Words (BoW)

Bag of Words is used in multiple aspects, it is utilized for information retrieval and analysis. The NLP techniques are used to extract a content to the data extracted in relativity to the objective of creating document classification [5]. The input to the bag of words model is as follows:

A dataset with 766 rows and 674 unique values, the vector space model is defined as follows: Suppose, the dataset has $\boldsymbol{N = 674}$ unique words $\boldsymbol{w_1, w_2, w_3, \dots\dots\dots w_{674}}$ from the dataframe $\boldsymbol{df}$, **e**ach document is characterized by an $\boldsymbol{N}$-dimensional vector whose $\boldsymbol{i^{th}}$component is the frequency of word $\boldsymbol{w_i}$ in the dataframe.

The steps involved in bag-of-words is the collection of data, creating design vocabulary and finally creating document vectors. The result of this is mentioned in the experimental analysis section. The disadvantage of Bag-of-words, is that it cannot retain the semantics, it can help to overall rank the number of occurrence of the words [13].

2.1.2 Term-Frequency Inverse document frequency (TF-IDF)

TF-IDF uses a methodology to quantify words from a set of words or set of documents. It compute a rank for every words to signify the meaning in the source files or input sequence or a corpus [7]. The intuition detection to rank the importance of the measure is the frequency of word appearance in the text sequence or document. The formula to rank the measure is:

$$tfidf\ (t, d|s, D|S) = tf(t, d|s) * idf\ (t, D|S) \dots\dots\dots\dots\dots\dots\dots$$

Where $\boldsymbol{t}$ represents the terms, $\boldsymbol{d \mid s}$ represents the data frame or sentence, $\boldsymbol{D|S}$ represents the collection of data frames or sequence of documents. TF-IDF gives the overview of the most unique words and the lowest ranked words can be considered as the most occurring in all the documents,

therefore these are the words the classifier will rely on for better leaning the similarity and linking.

2.1.3 Named Entity recognition (NER)

Ontology enrichment can be achieved on a new dataset by using natural language processing techniques such as Named entity recognition (NER), it performs entity chunking, extraction or identification. This will help in linking the documents to other similar documents. An entity can be a singular word or a sequence of words that consistently inherit similar meaning.

In this work, statistical method of entity extraction, to model the $\boldsymbol{probability(Y|X)}$ where Y is the sequence of named entity and X is the sequence of words.

Therefore, using discriminative model:

$$Probability\ (Y|X) = \frac{Probabilty(X|Y).Probability(X)}{Probability(Y)} \dots\dots\dots\dots\dots [13]$$

NER is efficient in finding semantics between the entity and the words sequence, the drawback or the challenge face by NER is detection of semantics for the words or sequence of words containing same or similar meaning [7].

2.1.4 Latent Semantic Analysis (LSA)

Latent semantic analysis is a method of extracting and representing meaning. LSA accounts to new data, and this is one of the most important aspect in this work, to identify and extract meaning out of the unstructured data. As unstructured data text extraction suffers with dimensionality [6]. LSA is a very effective method that can automatically extract the semantic similarities between the text sequence extracted from the document set , weight $\boldsymbol{w[n]}$ is filled corresponding to the rank of the terms in the source dataset and then finally reduced using singular value decomposition to reduce the dimensional space called concept space using the formula 4.

$$df = D1 \sum D2^T \dots\dots\dots\dots\dots\dots\dots [13]$$

Where $\boldsymbol{df}$ the data is frame of length 766, $\boldsymbol{D1}$, $\boldsymbol{D2^T}$ are the orthogonal parameter to reduce the dimension. After the dimensional reduction, the LSA then performs ranking $\boldsymbol{r}$ using the equation 5, approximating to the smaller concept space for $\boldsymbol{df} * \boldsymbol{df}$ matrix.

$$df_r = D1_r(\sum D2^T)_r \dots\dots\dots\dots\dots\dots [13]$$

This model is explored as it preserves relationships between words, which was seen as one of the disadvantages of bag-of-words NLP technique.

2.1.5 Latent Dirichlet Allocation (LDA)

In Natural Language Processing, LDA is a generative statistical model that allows a set of words to be linked to the groups that contain similarity. The usage of LDA can help reducing the dimensionality of the dataset, helps with class separation, avoids over-fitting and reduces the cost of computation. The probability model is calculated using the formula

$$Probability\ (R \mid \propto B) = \prod_{i=1}^{n} \int Probability(\emptyset_i \mid \propto) [\prod_{j=1}^{m(n)} Probability\ (W_{ij}|\emptyset_i, B)] d\emptyset_i \dots\dots [6]$$

Where $\boldsymbol{Probability\ (R \mid \propto B)}$ encodes the probability of the given topic in n-dimensional array.

In this paper, extracting text from the images using image processing is done, which is then store into the formats supported by the above natural language processing techniques [15], the background and its disadvantages give better understanding of the extent to which the meaning or semantics from the documents can be extracted.

III. INSIGHTS

The dataset is collected from the university library and collected dataset is to establish a systematic approach to structure the data to support the search analysis and output the most relevant document. The data collection method is standardized across all fields of study such as physical data collection, science, businesses etc. which the method of collection varies across different disciplines, here, data is collected from the college library, the library has the data stored in different file structure however there is no proper data labeling or structure which is the objective of this methodology. The dataset collected belongs to different classes such as circulars, notifications, fee structure, notes etc. These data should classify automatically to the respective departments.

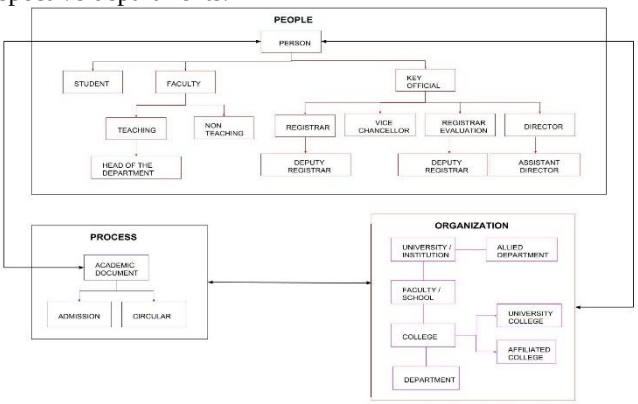

Fig.1 University data structure

IV. CONCLUSION AND FUTURE RESEARCH WORK

The paper provides a literature review on the concepts of Semantic Web classification using NLP techniques. Furthermore, the paper proposes a different NLP algorithms to classify the different sources of data.

Future work includes:

1. I shall classify the document among different departments. I should collect all the circulars, notifications, exam timetable, non-teaching staff circulars, sports department circulars etc, these files has to classify automatically by using different algorithms.
2. Implement the different types of natural language processing algorithms to generate the semantics from unstructured document.
3. Implementation of bag-of-words to count the similar words from the documents and TF-IDF techniques

allows the words to be ranked based on their occurrence, the lower the rank the higher the occurrence, this can be used to discard obvious English words.

4. Named entity recognition techniques allows to automatically analyze the association of the word or sequence of words to the already existing entities, this can give better grounds to create efficient semantics based on the entity definitions.

REFERENCES

[1] S. S. Rao and A. Nayak, “LinkED: A Novel Methodology for Publishing Linked Enterprise Data”, CIT. Journal of Computing and Information Technology, Vol. 25, No. 3, September 2017, 191–209 191 doi:10.20532/cit.2017.1003477

[2] C. Bizer , T. Heath, and T. Berners-lee, "Linked Data – The Story So Far", *International Journal on Semantic Web and Information Systems*, vol. 5, no. 3, pp. 1–22, 2009.

[3] Rodriguez, Danissa V.; Carver, Doris L.; Mahmoud, Anas (2018). [IEEE 2018 IEEE Aerospace Conference - Big Sky, MT, USA (2018.3.3-2018.3.10)] 2018 IEEE Aerospace Conference - An efficient wikipedia-based approach for better understanding of natural language text related to user requirements. , (), 1 16. doi:10.1109/AERO.2018.8396645

[4] G. Antoniou and F. Van Harmelen.” A Semantic Web Primer” MIT Press second edition, 2008.

[5] Singh, Sonit. (2018). Natural Language Processing for Information Extraction.

[6] Xie, Qingsheng; Zhou, Xiaoping; Wang, Jia; Gao, Xinao; Chen, Xi; Chun, Liu (2019). Matching Real-World Facilities to Building Information Modeling Data Using Natural Language Processing. IEEE Access, 7(), 119465–119475. doi:10.1109/ACCESS.2019.2937219

[7] Qaiser, Shahzad & Ali, Ramsha. (2018). Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents. International Journal of Computer Applications. 181. 10.5120/ijca2018917395.

[8] Ghannay, S.; Caubriere, A.; Esteve, Y.; Camelin, N.; Simonnet, E.; Laurent, A.; Morin, E. (2018). [IEEE 2018 IEEE Spoken Language Technology Workshop (SLT) - Athens, Greece (2018.12.18-2018.12.21)] 2018 IEEE Spoken Language Technology Workshop (SLT) - End-To-End Named Entity and Semantic Concept Extraction from Speech. , (), 692–699. doi:10.1109/SLT.2018.8639513

[9] Kim, Suhyeon; Park, Haecheong; Lee, Junghye (2020). Word2vec-based latent semantic analysis (W2V-LSA) for topic modeling: A study on blockchain technology trend analysis. Expert Systems with Applications, 152(), 113401–. doi:10.1016/j.eswa.2020.113401

[10] E. S. Negara, D. Triadi and R. Andryani, "Topic Modelling Twitter Data with Latent Dirichlet Allocation Method," 2019 International Conference on Electrical Engineering and Computer Science (ICECOS), 2019, pp. 386-390, doi: 10.1109/ICECOS47637.2019.8984523.

[11] Block, C., Wustmans, M., Laibach, N., & Bröring, S. (2021). Semantic bridging of patents and scientific publications – The case of an emerging sustainability-oriented technology. Technological Forecasting and Social Change, 167, 120689. doi:10.1016/j.techfore.2021.12068

[12] Sarica, S., & Luo, J. (2021). DESIGN KNOWLEDGE REPRESENTATION WITH TECHNOLOGY SEMANTIC NETWORK. Proceedings of the Design Society, 1, 1043-1052. doi:10.1017/pds.2021.104

[13] Prathyusha, K. S., & Reddy, B. E. (2021). Normalization Methods for Multiple Sources of Data. 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS). doi:10.1109/iciccs51141.2021.9432

[14] Gastaldi, Juan Luis (2020). Why Can Computers Understand Natural Language? Philosophy & Technology, (), –. doi:10.1007/s13347-020-00393-9

[15] Xue, F., Wu, L., & Lu, W. (2021). Semantic enrichment of building and city information models: A ten-year review. Advanced Engineering Informatics, 47, 101245. doi:10.1016/j.aei.2020.101245