INTELLIGENT RETRIEVAL
Open Journal

Review

Intelligent Information Retrieval Based on Semantics using Data Mining and Ontologies

Muqeem Ahmed*
Department of CS and IT, MANUU, Hyderabad, India

****Corresponding author***
Muqeem Ahmed
Department of CS and IT, MANUU, Hyderabad, India

Article information
Received: May 10th, 2023; **Revised:** June 3rd, 2023; **Accepted:** July 1st, 2023; **Published:** August 20th, 2023

Cite this article
Ahmed M. Intelligent information retrieval based on semantics using data mining and ontologies. 2023; 1(1). doi: https://doi.org/10.70705/ppp.ir.2023.v01.i01.pp37-41

ABSTRACT

The ever-increasing quantity of material saved and shared on the Web and other document repositories is causing the well-known challenges and issues associated with discovering and effectively managing information in large numbers. Over the last ten years, thanks to advancements in search engine technology The collection, storage, and pre-processing of information on a global scale has made remarkable strides, allowing for the immediate return of appropriate resources in response to user demands. Although the desired information may be there in the search area, users may still fail to find it or have to exert a lot of effort to get there. The current state of information retrieval methods relies on keyword-based consolidated content description and query processing, which limits their ability to understand and use user demands and content meanings. A large amount of research in the field of information retrieval has focused on the idea of concept based information retrieval, which involves solving the limitations of keyword based by searching or retrieving by meanings rather than literal strings or keywords. Conceptual search involves the inability to describe relations between search terms. The evolution of information retrieval based on semantics may be aided by semantic technologies like XML and ontology. Exploring the use of semantic technologies like XML and ontology to support more expressive queries and more accurate results, this paper aims to develop semantic based information retrieval to intelligently support semantic information retrieval search capabilities in large document repositories. In order to accomplish this, we gathered documents from various domains, organized them into tree structures using XML and ontology, and applied data mining techniques like clustering. Then, we used user interests to retrieve information from this structure, which provided the basis for the concept.

Keywords
Information retrieval; Ontology; Datamining; Semantic web.

INTRODUCTION

The quantity of information kept and made available online in various document repositories is growing at a rapid and constant rate. Some of the well-known challenges and concerns caused by this expansion include discovering and effectively managing the vast amounts of information that already exist. In the last ten years, search engine technologies have come a long way. These engines now gather, store, and pre-process user queries in order to return relevant resources that can help them achieve their goals, regardless of whether the query already exists in the search space or not. As a result of their reliance on keywords, the current state of information retrieval query processing and content description methods provide little capacity to understand and make use of conceptualizations related to user wants and the meanings of the material. The incapacity to adequately handle language problems like polismy or synonymy, as well as the difficulty in describing interactions between search words, are examples of these constraints. Information retrieval studies have found solutions via the use of novel conceptual information retrieval techniques.

Search, here meaning searching by meanings rather than literal strings [1], is one of the shortcomings of keyword-based information retrievals. Not all of these approaches take use of relevant semantic developments; others rely on quantifiable tactics that look at term co-occurrence. Terms' relationships are deduced from their frequency without taking into account possible issues like the polis my phenomenon. The degree of conceptualization is often weak and sparse, particularly at the relational level, in many other information retrieval systems that employ the language approach [2] that is analogous to the human mind. Although there are significant issues with conceptual information retrieval, the notion of providing a higher-level conceptual understanding of contents and queries has been prevalent in information retrieval since the early eighties [3]. After that, we touch on a few significant

developments in the field of information retrieval, such as the creation of automated methods for document analysis and Boolean query processing, as well as new probabilistic retrieval models. Initiated with the goal of facilitating the automation of processes requiring a presumption about the nature of the things at play or the goal of the endeavor, and the ability for computer programs to consistently discover, exchange, and combine relevant data and resources. At its heart, semantic web technologies are ontologies [5], which are seen as essential building blocks for representing conceptual information in a way that is understandable, usable, and shareable across distributed environments. This was an early attempt to address the shortcomings of traditional keyword-based information retrieval and search within the Information Retrieval context, and it has since been studied by various researchers. Standards for information modeling and codification were developed by the Semantic Web [6]. Intelligent information processing has become dependent on Semantic Web technologies like ontologies, XML, and data mining methods, which provide a structure for exchanging domain-specific conceptual information. User requests may be translated into a collection of discrete ideas using Web Ontology Language (OWL) [7] and onceptNet [8] for content searching and retrieval. The similarity of the documents is determined by merging the query words with semantically enhanced phrases found in both the texts and the clusters' labels and expand concepts. The original foundation of this semantic similarity method is WordNet, which has become the gold standard for describing ontologies on the Semantic Web [9]. Therefore, there is a chasm between data, applications, and data mining algorithms; this chasm is commonly called the semantic information retrieval research gap. When used in conjunction with data mining techniques, semantic web technology can intelligently fill in the gaps and extract the pertinent information. Semantic Web technologies, which formally describe domain information—including organized collections of prior information, inference rules, knowledge enhanced datasets, etc.—could therefore provide frameworks for the systematic integration of domain knowledge into an intelligent data mining environment. The user model [10] is derived from an analysis of user behavior in the system to capture user information based on user interests [11], and it is used to provide semantic information retrieval capabilities in web publications using Semantic Web technology and the WordNet ontology. However, the study mentioned above is still in its early stages. Improving semantic technology is necessary for faster concept-based information retrieval rather than keyword-based retrieval, which is necessary for handling both keyword-based and user interest-based information retrieval. This work presents an effort to create an information retrieval interface that relies on clustering, XML, and ontology matching, and is focused on concepts and user interests. A large number of resource descriptions use diverse terminology to express the same or similar ideas inside multiple ontologies; this is due to the distributive nature of the Semantic Web. It finds matches for those parts, and then, using more detailed queries, it returns more accurate results to get the discovering knowledge using semantic analysis. Thus, by using semantic technology and data mining methods, the suggested framework for concept-based information retrieval will manage information based on concepts and user interests to a greater extent.

II. RELATED WORK

The advances in learning building and data mining advance semantic information mining, which conveys rich semantics to all phases of information mining process. Many research endeavors have bore witness to the benefit of fusing area learning into information mining. Formal semantics encoded in the cosmology is all around organized which is simple for the machine to peruse and process in this manner make it a nature approach to utilize ontologies in semantic information mining. Utilizing ontologies, semantic information mining has points of interest to connect semantic holes between the information, applications, information mining calculations, and information mining comes about, give the information mining calculation with priori learning which either manages the mining procedure or lessens the pursuit space, and to give a formal approach to speaking to the information mining stream, from information pre processing to mining comes about Information retrieval called learning disclosure from database, is the procedure of nontrivial extraction of understood, already obscure, and possibly helpful data from information [12]. In the previous couple of decades, propels in Data mining systems prompt numerous momentous transformations in information investigation and huge information. Data mining likewise joins methods from insights, manmade brainpower, machine learning, database framework, and numerous different controls to investigate expansive informational collections. Semantic Data Mining alludes to information mining errands that efficiently join area learning, particularly formal semantics, into the procedure. The viability of space information in information mining has been bore witness to in past research endeavors.

Fayyad et al. [13] guaranteed that space information can assume a vital part in all phases of information mining including, information change, highlight decrease, calculation choice, post-preparing, display translation et cetera. Russell and Norvig [15] trusted that an insightful operator (e.g., an information mining framework) must be able to acquire the foundation information and ought to learn learning all the more successfully with the foundation learning. Past semantic information mining research has confirmed the positive impact of area learning on information mining. For instance, the pre processing can profit by space learning that can enable channel to out the repetitive or conflicting information [14], [16]. Amid the looking and example creating process, area learning can fill in as an arrangement of earlier information of requirements to help decrease seek space and guide the inquiry way [17], [18]. Encourage more, the found examples can be successfully removed
they may be encoded in the formal structure of information design [21] or made more evident [19], [20]. First things first when using space learning for data mining: make sure your PC can access and analyze the models you need for communicating and building your learning. A series of area learning techniques that officially manufacture and use space information has been pleasantly enhanced by the growth of information designing (KE) [22]. The explicit determination of a conceptualization is one of the beneficial advancements in information creation that is metaphysics [23], [24]. It is common practice to construct a cosmology to denote a certain domain (for instance, inherited traits). Formally indicating the concepts and relationships in that space, this kind of metaphysics is sometimes re-

ferred to as an area cosmology. Successful knowledge exchange and reuse relies heavily on the embedded formal semantics found in ontologies. Gene Ontology (GO) [25], Unified Medical Language System (UMLS) [26], and over 300 ontologies in the National Center for Biomedical Ontology (NCBO) [27] are notable examples of space ontologies. Very developed ideas for showing and categorizing area learning have been suggested by research in the Semantic Web zone [28]. Semantic Web ontologies have emerged as a game-changing development in smart data preparation, facilitating the exchange of domain-specific applied models. A Language for Web Ontologies (OWL)

For this reason, most people use [29], which has become the gold standard for describing ontologies used on the Semantic Web. Structures for orderly combining of area information in an astute information mining condition might be created by the Semantic Web innovations that explicitly speak to space learning, such as structured accumulation of previous data, surmising rules, information improved datasets, and so on. Our focus in this review article is on the current state and future prospects of semantic information mining as a field of study. Methodologies grounded in philosophy are our primary focus. Semantic information mining approaches grounded in cosmology aim to include formal ontologies into the data mining process. Most of the time, this is achieved by controlling the information mining process with the aid of ontologies, which provide formal meaning to concepts and relationships. To illustrate the consistent connections of the arranging attempt, metaphysics might be used in categorization, for example. The grouping project would lead to better accuracy by deciding out the competing search space [30]. The information mining outcome may be adequately represented by an orderly association of ontologies, therefore there's no need to discourage this. In data extraction and content mining, for example, the philosophy itself may be used to display the separated data using a cosmological specification dialect (like OWL) [31]. Over the last ten years, there has been much discussion within the information mining community about how to manage and govern the massive amounts of data. Because of the enhancement

An important area of study for the group is the construction of information, with a focus on Semantic Web technologies, mining large amounts of semantically rich and heterogeneous data. While a similar amount of researchers have raised the topic, semantic information mining remains in its infancy. One of the most promising approaches is philosophy-based semantic information mining. The true test is in developing more automated semantic information mining algorithms and structures by using the completeness of formal philosophy, which has greatly defined representational dialect, formal semantics, and reasoning tools for consistency testing and induction.

METHODOLOGY

Figure1: Semantic based information Retrieval

Initially, text documents which have been collected from various sources were accumulated in a database. Then, pre- processing was carried out by considering the various stages like tagging by means of Stanford POS tagger tool, stop word removal and stemming, based on Porter Stemmer algorithm and morphological capabilities of WordNet. The above pre processing is common for existing algorithm. The proposed algorithms considered in this study are based on Modified K- Mean clustering. It represented the tree database of collected learning documents. These documents are clustered using modified K-means algorithm which generates K-Number of clusters and the algorithms as originally proposed by the various authors were implemented in the above environment. For uniformity the clustering- based concept semantic similarity alone is implemented with query retrieval and then proceeds to clustering. By varying the number of documents the results of the proposed and existing algorithms are measured. These algorithms are implemented in JDK 1.7 environment using Net Beans IDE

III. RESULTS AND DISCUSSION

Each item's average maximal possibly frequent set size and average transaction size are both set to maximum. Using the modified approach method yields different cluster outcomes compared to the traditional K-means algorithm. The time it takes to execute the record is higher in the k-means algorithm compared to the modified k-means algorithm because the heterogeneity of group balancing is an issue. So, to get around this difficulty and keep both groups balanced, we adopted the spasis approach. Therefore, compared to the K-Means Algorithm, the time required to execute the records is lower. The method's first step is identical to that of the k-means algorithm; when we receive the final result, we adjust the clustered data according to the clusters so that they're evenly distributed.

The comparison between K-mean and Modified approach K- mean on the basis of large number of records and execution time using this algorithm showed in the above figure. The Modified approach K-mean is better performance comparison to standard K-means algorithm.

the xml and ontology representation of the data that focus information retrieval more intelligently comparison to the previous approach of the information retrieval such as keyword based approach. This approach introduced concept based information by utilizing the semantic technologies. It can also provide better search capability to achieve some qualititative improvement over the keyword based information retrieval by introducing the xml and domain ontology. For the implementation of this we have collected hundreds documents from the different domain and designed the tree data structure in the form of xml and ontology by using java and protégé ontology editor and get the concept based results from the collected documents . As future research line we plan to explore other Semantic Web gateways, to make use of larger amounts of online available semantic metadata. Another important research problem, aside from obtaining high quality knowledge resources, is the annotation of unstructured content. The annotation problem consists in identifying semantic entities within the contents. It is a difficult research problem on its own, which is being widely studied in areas such as Information Retrieval, NLP and Semantic Web.

IV. CONCLUSION AND FUTURE SCOPE

A relatively new innovation in the realm of information retrieval and the semantic web, semantic information retrieval emerged as a result of the expansion of information technology. The primary

goals of this platform are to improve the organization of data and to make it easier to find and retrieve relevant information inside the information space. The aforementioned literature study also notes that the current method has several drawbacks, such as the fact that semantic information retrieval is slow to apply due to factors like the amount of the data and the variety of web-based information. Intelligent information retrieval based on semantic web technologies combines traditional methods of retrieval with semantic web technologies such

REFERENCES

(1) Susan T. Dumais. Furnas, George W. Landaue, Thomas K."Latent Semantic Analysis Indexing" Bell Research on Communications 1990

[2] in"Sense clusters for information retrieval" by Gonzalo, Verdejo, Chugur, and Cigarrán The COLING/ACL'98 Workshop on WordNet for Natural Language Processing, Montreal, 1998

the thirdAccording to Bruce Croft's 1986 paper "Boolean queries and term dependencies in probabilistic retrieval,"

"Information Retrieval" by C. J. van, published in the ACM SIGAR Forum, volume 17, issue

[4], 1979.

[5] "A Translation Approach to Portable Ontology Specifications" by Thomas R. Gruber. "Knowledge Acquisition," volume 5, issue 2, pages 199–220, 1993.

In their article "The Semantic Web," T. Berners-Lee, J. Hendler, and O. Lassila discuss a topic. November 2001, Scientific American, 284(5): 28–37

7] Web Ontology Language (OWL). Please visit http://www.w3.org/TR/owl-ref/.

[8] "Commonsense Knowledge Extraction Using Concepts" Properties by Blanco, Cankaya, and Moldovan. An International Conference on Artificial Intelligence, Twenty-Fourth Edition, 2011.

The paper "Comparison of Current Semantic Similarity Methods in WordNet" was written by Li, H., Tian, Y., Ye, B., and Cai, Q. Tenth Annual International Conference on Computer Application and System Modeling (ICCASM 2010). 10. IEEE, 978-1-4244-7237-6.

In their paper titled "User Profile Modeling and Learning," Nedelkou et al. (2010) cite a number of sources. Included in the "Encyclopedia of Information Science and Technology," Second Edition. Chapter 627, pages 3934–3939, has the DOI 10.4018/978-1-60566–026-2. We are IGI Global.

A Semantic Web-Based Approach to Adaptive E-Learning: Learning and Learner Profile Update

[11] Harb, H., & Fouad, K. The Eleventh International Conference on Al-Azhar Engineering, December 23–26, 2010

[12] "Data Mining: Concepts and Techniques" by J. Han and M. Kamber. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2005.

"From data mining to knowledge discovery in databases"

[13] is written by U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. Journal of artificial intelligence, 17(3), 37, 1996.

[14] "Active user-based and ontology-based web log data preprocessing for web usage mining" (Khasawneh and Chan, 2014). Web Intelligence 2006: Proceedings of the IEEE/WIC/ACM International Conference, pp 325-328, 2006.

[15] "Artificial Intelligence: A Modern Approach." S. J. Russell and P. Norvig. Third edition, 2003, Pearson Education.

[16] "Ontodataclean: Ontology-based integration and preprocessing of distributed data" by D. Perez-Rey, A. Anguita, and J. Crespo. The book "Biological and Medical Data Analysis" includes chapters 262-272. Press, 2006.

Referenced by Y. Mansour, A. Blum, and N. Balcan [17]."Learning by utilizing ontology structures and unlabeled data." Chapters 1112–1120 of the 2013 Proceedings of the 30th International Conference on Machine Learning.

The authors of the paper "Ontology-driven association rule extraction: A case study" are A. Bellandi, B. Furletti, V. Grossi, and A. Romei. Representation and Reasoning with Contexts and Ontologies, 2007, page 10.

The authors of the article "Knowledge-based interactive postmining of association rules using ontologies" are C. Marinica and F. Guillet. Journal of Electrical and Computer Engineering, Volume 22, Issue 6, Pages 784–797, 2010.

[20] "Using ontologies to facilitate post-processing of association rules by domain experts" (G. Mansingh, K.-M. Osei-Bryson, and H. Reichgelt). Science of Information, 181(3):419-434, 2011.

"Components for information extraction: Ontology-based information extractors and generic platforms" (D. C. Wimalasuriya and D. Dou, 2021). The 19th Annual Conference on Information and Knowledge Management (CIKM) Proceedings, 2010, pages 9–18.

AI: A Contemporary Perspective, by S. J. Russell and P. Norvig

[22]. Third edition, 2003, Pearson Education.

[23] In the International Journal of Human-Computer Studies, T. R. Gruber wrote about "Toward principles for the design of

ontologies used for knowledge sharing" in 1995 (vol. 43, no. 5, pp. 907-928).

The authors of "Knowledge Engineering: Principles and Methods" are R. Studer, V. R. Benjamins, and D. Fensel. Volume 25, Issue 1, Pages 161–197, 1998, in the journal Data & Knowledge Engineering.

The gene ontology consortium
[25] references. "Design and implementation of the gene ontology resource: creating." August 2001, Genome Research, 11(8):1425–1433.

"The Unified Medical Language System"
[26] by D. Lindberg, B. Humphries, and A. McCray. Article published in 1993 in the journal Methods of Information in Medicine, volume 32, issue 4, pages 281-291.

This information is sourced from the National Center for Biomedical Ontology. Visit http://www.bioontology.org/./.

In their 2001 article "The Semantic Web," T. Berners-Lee, J. Hendler, and O. Lassila discuss the topic in Scientific American in its 284(5):28-37.

The Open Web Ontology Language (OWL)([29]). Please visit http://www.w3.org/TR/owl-ref/.

[30] Yusof Mansour, A. Blum, and N. Balcan.Using ontology structures and unlabeled data to facilitate learning. Chapters 1112–1120 of the 2013 Proceedings of the 30th International Conference on Machine Learning.

"Hybrid ontology-based information extraction for automated text grading" [31] (F. Gutierrez, D. Dou, A. Martini, S. Fickas, and H. Zong). This is volume 1 of the 2013 12th International Conference on Machine Learning and Applications (ICMLA), which spans pages 359 to 364. 2013, IEEE.

(32, 31) Authors D. C. Wimalasuriya and D. Dou provide "Ontology-based information extraction:" an overview of the field and existing methods. The 2010 Journal of Information Science, volume 36, issues 3, pages 306-323.