SEMANTIC WEB AND ONTOLOGIES

Adwait Desai [1] Mr. Sandip Warhade [2]

[1]Student; Pune Institute of Computer Technology, (department-IT), Pune, Maharashtra, India, adwait393@gmail.com
[2]Assistant Prof.; Pune Institute of Computer Technology (department-IT), Pune, Maharashtra, India, srwarhade@pict.edu

Abstract:

The Semantic Web, the cornerstone of web 3.0, relies on ontologies to manage data heterogeneity and enable automated information repurposing and analysis. However, choosing an appropriate ontology in line with user requirements remains a challenging problem due to the time and effort required, the lack of context awareness, and the computational complexity. This work proposes an ontology recommendation system that combines text classification with unsupervised learning techniques to overcome these challenges. The proposed study offers a number of benefits, including minimal computational complexity, effective ontology organization, reduced time and effort spent selecting appropriate ontologies, and adaptability to a range of domains and online ontology libraries. At last, I have provided the case study for better understanding how a complete semantic web will work.

Keywords: recommendation system, data heterogeneity, unsupervised learning, ontologies, text categorization, Semantic Web

1.Introduction

The topic of this seminar proposes an ontology recommendation system that combines text classification and unsupervised learning techniques to suggest the optimal ontology based on user requirements, grouping ontologies according to comments from domain experts. The article also includes a description of the software requirements for the proposed framework. It also discusses a variety of approaches and various Machine Learning Algorithms, such as K-Means Hierarchical clustering. Basically, I have compared the three methods to understand which one is the most suitable for semantic web and ontologies.

Table 1.1 Comparison between Semantic and Traditional Web

Characteristic	Semantic Web	Traditional Web
Data Representation	RDF, structured data	HTML documents, unstructured
Data Meaning	Machine-readable, annotated data	Primarily human-readable text
Data Interconnection	Strong interconnection through RDF triples and linked data	Limited interconnectivity between documents
Search and Discovery	Semantic search engines, context-aware	Search engines based on keyword matching
Human Interpretation	Automated processing, reasoning	Relies on human interpretation and browsing
Data Integration	Automatic integration through ontologies and RDF graphs	Manual integration and data transformation
Data Inference	Supports inference, reasoning	Lacks inference capabilities
Scalability	Scales well due to structured data	Becomes challenging with unstructured data growth
Data Consistency	Promotes data consistency and integration	Inconsistent and prone to data silos
Domain Knowledge	Encourages domain-specific ontologies	Lacks formalized domain knowledge
Semantic Interoperability	Enhanced semantic interoperability through shared ontologies	Limited semantic interoperability
Data Trustworthiness	Promotes data trustworthiness and provenance	Data source reliability can be challenging to determine
Real-World Applications	Used in knowledge graphs, data integration, intelligent agents, etc.	Primarily used for information sharing and e-commerce

2.Literature Survey

Table 2.1 Literature Survey

Year	Author	Paper Name	Description
2020	Mohsin Raza,Mansour Ahmed,Asad Habib	Exploting Ontology using text categorization	Provides overview of Ontology Recommendation and text categorization approach, compares different ML algorithms to find out the best use of each.
2023	Weijun Tan	Overlooked video classification inWeakly supervised anomaly detection	Basically this paper is about deep learning approaches which involve video classification.
2023	Yijin Lin,Zhipeng Gao,Hongyang Du,Dusit Niyato	A unified Blockchain-Semantic Framework for Wireless Edge Intelligence Enabled Web 3.0	Provides a framework which integrates blockchain technology, semantic web and wireless edge computing address the challenges of Web 3.0
2023	Xu Zhang, Tong Li, Zhan Ma	AI and Blockchain Empowered Metaverse for Web 3.0: Vision, Architecture and Future Directions	This paper provides the architecture for AIB-Metaverse , which brings in the picture together usage of AI and Blockchain.

3 .Proposed Methods

2.1.1 Proposed Framework

The proposed design calls for the establishment of an ontology repository, the gathering of user needs, and the construction of an unsupervised learning and text classification-based recommendation system. Based on user needs, the framework offers the most relevant ontology by grouping similar ontologies into clusters. By proposing the one most suitable ontology, this framework seeks to get over the drawbacks of giving users a plethora of results. To save developers time and effort, it uses unsupervised learners and text classification to suggest the best ontology to the user.

By supporting data providers, information engineers, and ontology designers—both fresh and experienced—identify the right ontology and cut down on the time and effort needed to do so, the suggested approach may be used to promote ontologies. In order to analyze how well the ontology recommendation engine organizes ontologies, predicts the appropriate ontology group, and recommends ontologies based on user needs, the framework also has a performance assessment model .

Four stages constitute the building process of the framework's functionality: ontology crawling, pre-processing tasks, unsupervised learning, and ontology suggestion. In ontology crawling, ontology words and text are

obtained, pre-processing operations are carried out over user needs and ontology data, related ontologies are grouped using unsupervised learning, and an ontology is recommended for the specified user demand .

A). Ontology Recognition:

The methodical process of collecting ontologies from many sources, including literature, internet databases, and domain-specific databases, is known as "ontology recognition." This entails locating, obtaining, and compiling ontologies relatable to certain domains or topic areas. Establishing and upholding ontology repositories requires ontology crawling, which makes sure that a wide variety of ontologies are available for additional examination and advice. Concerning the topic under discussion, ontology recognition plays a pivotal role in enriching the ontology collection with a variety of relevant and varied ontologies from various domains. This allows the framework to suggest the most suitable ontology to users depending on their specific requirements.

B). Pre-Processing

Preprocessing includes a series of procedures to get the data ready for further examination. I've covered word indexing, lemmatization, and stop-word elimination as three methods for pre-processing here. To decrease data sparsity and feature set size, stop-word removal removes words like propositions and pronouns that convey no meaning or information. Lemmatization reduces word duplication caused by capital or lowercase variations by grouping inflected variants of a word into a single item.

Stop-word Removal: Removing words that give no relevance to the data which may be adjectives.
Lemmatization: Uniform representation of words to remove data duplication.
Word Indexing: Texts are converted to numeric data to make data more analysable.

C) Clustering

Cluster is collection of related object, example is how we say cluster of stars. Clustering includes the groups of data formed for their analysis and to drive some conclusions or make predictions. It is a prediction based business intelligence method.

There are several clustering methods:
1)K-means
2)Hierarchical

1) K-means: Heuristic method, where each cluster is represented by center of clusters.
   K stands for umber of clusters.
   Algorithm:
   - Selection initial centroid at random.
   - Assign each object to cluster with nearest centroid.
   - Compute each centroid as the mean of objects assigned to it.
   - Repeat the steps until no change.

2) Hierarchical: Uses distance matrix as clustering criteria. This method does not require the number of clusters K as an input, but needs termination condition. Represented by a Dendogram . Logic used here is linkage.

Table 3.1.C- Comparison Of various Hierarchical Clustering approaches

Agglomerative Clustering	Divisive Clustering
Follows a top-down approach	Follows bottom up approach
Clustering occurs group by group as initially each item is in its own clusters.	First a large cluster is formed and based on similarities small clusters are made.

Some methods that are used to increase the accuracy of text categorization are as follows:

1. Binary: This basic weighting technique assigns a word a weight of 1 if it appears in the document and 0 otherwise. This technique is employed to indicate if a phrase is in a document or not .

2. Term Frequency-Inverse Document Frequency (TFIDF): This numerical metric illustrates the importance of a term to a document in a collection or corpus. In order to account for terms that appear more frequently overall, it considers both the frequency of a term in a document and the total number of documents in the corpus that contain the phrase.

3. Entropy: -Entropy is comparable to a randomness or surprise metric. It helps us comprehend how much a word may disclose about a particular document in the context of text classification.

4. Term Frequency Collection (TFC): TFC is a sophisticated variant of TFIDF, a text analysis tool. Unlike TFIDF, TFC examines the whole length of the document in addition to the frequency of terms appearing in it.

5. Length Term Collection: LTC is a clever method of managing the frequency with which words appear in a manuscript. To ensure that words which appear frequently enough are noticed and that ones that appear infrequently enough are not overemphasized, it employs a method involving logarithms.

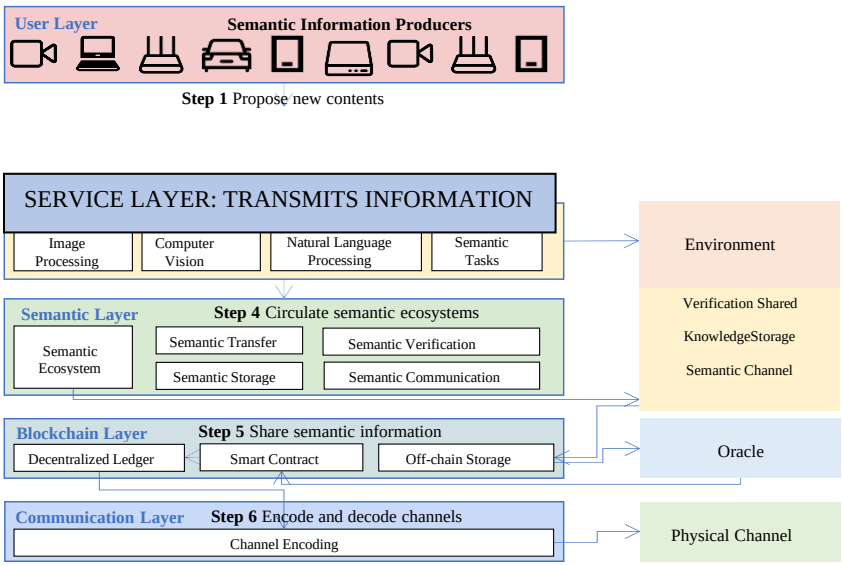

Figure 1: Probable Semantic Web Architecture ([1]Unified Blockchain-Semantic Framework)

4. Results and Discussions

Ontology need arises in many cases such as University databases. They can be used while choosing an engineering course as well.
Example: If you are confused about choosing a course, the counsellor asks you to name the subjects you are interested in. You say mathematics, physics, c-programming and Java. With the answers that you provided to the counsellor, he relates all different subjects and finds that the most suitable course for you is Computer Science course.
I have also explained a case study at the end, stating what makes the website completely semantic.

4.1.1 Dataset Discussion

I have taken the dataset from my reference paper [2] . The dataset in the paper is collected manually which consists of 30 user requirements. The evaluation is done, and effectiveness of the method is found out.
Why User Needs Are Important: - Assume You Have Queries Imagine that you are a person with a ton of queries, such as the desire to learn computer science ideas, discover the ideal recipe, or research academic subjects. Every one of these inquiries resembles a distinct demand you may have.

The Large Electronic Library: - Where Solutions Are Stored: Imagine a vast digital library with shelves crammed with knowledge on anything from computer science to academics, as well as recipes for your favorite foods and drinks.

Your Individual Assistant - The Structure: -What Functions It Has: Let's say you have a very intelligent digital companion that we'll refer to as the Framework. This acquaintance is proficient at using the online library. When you explain to your friend what it is you're interested in learning about (your user needs), they say, "Okay, let me locate the ideal section."

An ontology may be necessary in the academic area in order to administer university information systems, according to user needs. Criteria for separating and organizing data on classes, instructors, students, investigations, and academic departments may fall under this category. The ontology could be needed by users to make things like course scheduling, student enrolment, faculty administration, and research cooperation easier. Furthermore, in order to facilitate the effective integration and retrieval of data from diverse university databases and systems, the ontology might also need to record the relationships and characteristics of academic entities.

User needs in the science of computing domain could include the necessity for an ontology to describe ideas relevant to computer science, such as problem monitoring, methods for developing software, and software design patterns. To assist activities like programming, bug monitoring, and software architecture design, users might need the ontology. In order to facilitate the structuring and recovery of information pertaining to computer science ideas, the ontology might require to include the connections and qualities associated with software sections, design conventions, and development processes.

4.1.2 Comparison for different approaches that can be used to make the web Semantic:

Table 4.1.2.1 Comparing different algorithms

Characteristic	K-Means	K-Medoids / Content Recommendation
Algorithm Type	Clustering	Clustering / Recommendation
Purpose	Data clustering and segmentation	Data clustering and segmentation / Suggesting relevant content to users
Semantic Web Integration	Limited	Limited / Integral for personalized content recommendations
Data Types	Numeric data	Numeric data / Structured and unstructured data
Semantic Data Usage	Minimal	Minimal / Relies on semantic data for user preferences and context
Ontology Usage	Rarely involves ontologies	Rarely involves ontologies / May use domain-specific ontologies
User Engagement	Typically not focused on user interaction	Typically not focused on user interaction / Designed for user engagement
Real-World Applications	Data segmentation, customer segmentation	Data segmentation, outlier detection / Recommending products, articles, videos, etc.
Challenges	Limited use of semantic data	Limited use of semantic data / Handling semantic data and user preferences

4.1.3 Diagrams

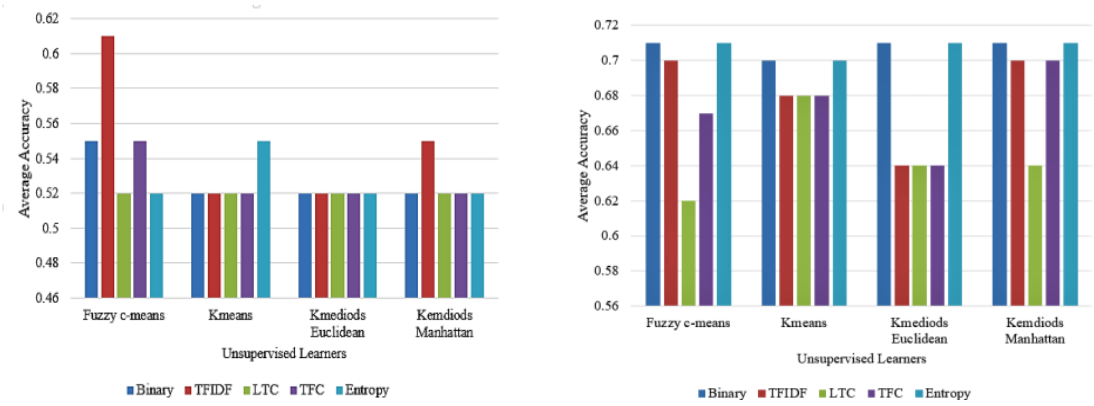

Figure 4.2.2 a) Academic domain b)Computer Domain ([2]Exploiting Ontologies)

	Fuzzy-c	K-means	Kmedoids	Kmedoids Manhattan
Academics	0.65	0.43	0.61	0.81
Computer	0.78	0.66	0.69	0.60

4.1.4 Conclusions

Inference from the Table:

Even though each algorithm has its own advantages and disadvantages, we cannot say which one is the best suitable algorithm to use in semantic web. The approach of using different algorithms can be determined with respect to the size of dataset provided or obtained. As clustering algorithms come in the category of prediction based algorithms, we cannot use evaluation metrices such as accuracy score as we use in description based algorithms (Eg:Naïve Bayes).

Now considering the example of dataset which is related to semantic web, where each data point represents a research paper. Research paper also includes metadata such as author information , abstract , keywords etc.
In this case if the dataset is too large, K-means will be used due to its computational complexity.
If the dataset contains topics, various subtopics which form a hierarchy then Hierarchical clustering will be used.

5.Acknowledgement

I would like to thank my institution Pune Institute of Computer Technology for providing such an opportunity and all the associated faculties who guided me to write my first ever paper.

6.References:

[1] Yijing Lin, Zhipeng Gao, Hongyang Du, Dusit Niyato, Jiawen Kang, Ruilong Deng, and Xuemin Sherman Shen. The paper is titled "A Unified Blockchain-Semantic Framework for Wireless Edge Intelligence Enabled Web 3.0".Y. Lin, Z. Gao, H. Du, D. Niyato, J. Kang, R. Deng, and X. S. Shen, "A Unified Blockchain-Semantic Framework for Wireless Edge Intelligence Enabled Web 3.0," in IEEE Transactions on Network Science and Engineering, vol. 9, no. 5, pp. 7650-7658, Sept.-Oct. 20

[2] M. A. Sarwar, M. Ahmed, A. Habib, M. Khalid, M. A. Ali, M. Raza, S. Hussain, and G. Ahmed, "Exploiting Ontology Recommendation Using Text Categorization Approach," in IEEE Access, vol. 9, pp. 27304-27315, 2021.

[3] "A Fair and Efficient Blockchain-Based Semantic Exchange Framework for Participatory Economy" and the authors are Hongyang Du, Jiawen Kang, Hui Yang, Dusit Niyato, Yaofeng Tu, and Zhipeng Gao.

[4] D. Kılıç, A. Özçift, F. Bozyigit, P. Yıldırım, F. Yücalar, and E. Borandag, ''TTC-3600: A new benchmark dataset for turkish text categorization,'' *J. Inf. Sci.*, vol. 43, no. 2, pp. 174–185, Apr. 2017.

[5] M. Javed, B. Ahmad, S. Hussain, and S. Ahmad, ''Mapping the best practices of XP and project management: Well defined approach for projectmanager,'' *J. Comput.*, vol. 2, no. 3, pp. 2151–9617, 2010.

[6] T. Korenius, J. Laurikkala, K. Järvelin, and M. Juhola, ''Stemming and lemmatization in the clustering of finnish text documents,'' in *Proc. 13th ACM Conf. Inf. Knowl. Manage. (CIKM)*, 2004, p. 625.

[7] M. Allahyari *et al.*, ''A brief survey of text mining: Classification, clus- tering and extraction techniques,'' Jul. 2017, *arXiv:1707.02919*. [Online].

[8] M. Lan, C. Lim Tan, J. Su, and Y. Lu, ''Supervised and traditional term weighting methods for automatic text categorization,'' *IEEE Trans. PatternAnal. Mach. Intell.*, vol. 31, no. 4, pp. 721–735, Apr. 2009.

[9] G. Salton and C. Buckley, ''Term-weighting approaches in automatic textretrieval,'' *Inf. Process. Manage.*, vol. 24, no. 5, pp. 513–523, Jan. 1988.

[10] T. Wang, Y. Cai, H.-F. Leung, Z. Cai, and H. Min, ''Entropy-based term weighting schemes for text categorization in VSM,'' in *Proc. IEEE 27th Int. Conf. Tools with Artif. Intell. (ICTAI)*, Nov. 2015, pp. 325–332.

[11] C. Zhang, X. Wu, Z. Niu, and W. Ding, ''Authorship identification from unstructured texts,'' *Knowl.-Based Syst.*, vol. 66, pp. 99–111, Aug. 2014.

[12] E. Saraç and S. A. Özel, ''An ant colony optimization based featureselection for Web page classification,'' *Sci. World J.*, vol. 2014, pp. 1–16, 2014.

[13] J. Du, W. Cheng, G. Lu, H. Cao, X. Chu, Z. Zhang, and J. Wang, "Resource pricing and allocation in mec enabledblockchain systems: An a3c deep reinforcement learning approach," *IEEE Transactions on Network Science and Engineering*, vol. 9, no. 1, pp. 33–44, 2021.

[14] A. Beniiche, S. Rostami, and M. Maier, "Society 5.0: Internet as if People Mattered," *IEEE Wireless Commu- nications*, 2022.

[15] Z. Liu, Y. Xiang, J. Shi, P. Gao, H. Wang, X. Xiao,B. Wen, Q. Li, and Y.-C. Hu, "Make Web3. 0 Connected," *IEEE Transactions on Dependable and Secure Computing*, 2021.