Medicon Agriculture & Environmental Sciences
Volume 7 Issue 3 September 2024
Article Type: Research Article
ISSN: 2972-2691

SOTW: Semantics Oriented Tagging of Web Pages

Akshith Gunasheelan[1]* and Gerard Deepak[2]

[1]*Compute Cloud Services Engineering, Hewlett Packard Enterprise*

[2]*Department of Computer Science and Engineering, Manipal Institute of Technology Bengaluru, Manipal Academy of Higher Education, Manipal, India*

***Corresponding Author:** Akshith Gunasheelan, Compute Cloud Services Engineering, Hewlett Packard Enterprise.

Received: August 03, 2024; **Publised:** September 11, 2024

Abstract

There is a need for a strategic semantically inclined model for tagging web pages in the era of the latest Web 3.0. This paper proposes a strategic semantic oriented knowledge driven learning infused model for tagging of web pages which encompasses extraction of terms and categories of web pages and application of models like TF-IDF and Structural Topical Modelling (STM) which is in-turn followed by generating the RDF instances and enriching the entities that come out of this pipeline through the Wikidata API. The proposed framework also uses a Logistic Regression Classifier Unit (LRU) which encompasses the RDF subject and object instances as the features to classify the web page dataset. A dynamic knowledge stack generation via a semantic agent and classification of the dynamically generated knowledge stack using a strong deep learning CNN classifier helps in increasing the overall learning capability of the model. Semantics oriented reasoning is achieved using the Adaptive Pointwise Mutual Information (APMI) measure with differential step deviance measures and the shuffled frog-leap algorithm takes care of the meta-heuristic optimization by improving the intermediate optimization results and overall precision of 95.18%, with a False Discovery Rate (FDR) of 0.05 and an F-measure of 96.30% which makes it the best in class model when compared to the other baseline models for semantics-oriented learning through webpage tagging.

Keywords: Normalized Pointwise Mutual Information (NPMI); Pointwise Mutual Information (PMI); Adaptive Pointwise Mutual Information (APMI); Logical Regression Unit (LRU)

Introduction

The World Wide Web (WWW) is a global system of digital information accessible via the Internet. It resembles a virtual society with diverse content, interconnected hyperlinks, virtual communities, and information exchange. Created by Tim Berners-Lee in 1989, it has transformed communication, learning, business, and global interactions. Web pages are crucial in the World Wide Web, employing languages such as HTML and CSS to organize and present various content. They support easy navigation through hyperlinks, contributing to the web's interconnected nature. Web 3.0, often coined as the "Semantic Web," signifies a progression in the evolution of the internet. It aspires to transcend the current state of the World Wide Web by introducing a more sophisticated and interconnected digital realm, focusing on both human and machine comprehension of information. This vision entails structuring data not only for human consumption but also in a manner that computers can grasp and interpret its significance. This transition is fueled by technologies like linked data, ontologies, and machine learning, which empower advanced search capabilities, automated data integration, and context-aware applications.

The Semantic Web integrates metadata and contextual insights, enabling machines to understand complex data relationships through RDF and OWL. It empowers applications to provide tailored user experiences, like intelligent search engines and data aggregation systems. This heralds a paradigm shifts towards a more intelligent and intuitive internet landscape. Web page tagging is crucial, improving accessibility, organization, and searchability with descriptive labels and metadata. It fosters connections between pages, supports content syndication, and enhances personalized user experiences. A semantic, knowledge-centric paradigm for web page tagging is needed to boost usability and effectiveness.

Motivation: A semantic web page tagging framework is needed to align with the web 3.0 standard, addressing the limitations of existing non-semantic models. The proposed framework utilizes AgentSpeak and deep learning CNN models to generate a knowledge stack. It incorporates the APMI measure for semantics-based reasoning and the shuffled frog-leap algorithm for optimization, achieving a precision of 95.18%, FDR of 0.05, and an F-measure of 96.30%. This positions it as a leading model for semantics-oriented web page tagging.

Contribution: This research work formulates the following primary contributions for achieving a strategic model for semantics-oriented tagging of web pages. The inclusion of terms and categories for webpages through the TF-IDF and the structural topical modelling (STM) which is in-turn used to generate the triadic RDF entities in which only the subject and the object of the RDF is retained. These retained entities are retained for feeding it as features for the Logical Regression Classifier Unit (LRU) to classify the dataset of webpages and encompassing it with the Wikidata API and frame-work for auxiliary knowledge enrichment.

Organization: The subsequent sections of this document are arranged as follows: Section 2 presents an overview of Related Literature. In Section 3, the Proposed System Architecture is elucidated. Section 4 presents the Performance Assessment and corresponding outcomes, while conclusive insights are offered in Section 5.

Related Works

Sharma et al. [1] address challenges in contemporary internet search methods that often result in imprecise and overwhelming outcomes. They propose an innovative semantic information retrieval architecture with a carefully crafted algorithm to compute document ranks, improving search result relevance. The algorithm considers keyword frequency within web pages and the nuanced association with contextually meaningful terms. Rigorous testing confirms the algorithm's robustness in consistently delivering pertinent and contextually fitting results. In their study, Lu et al [2] introduce a content-based approach to tag recommendation, targeting webpages with or without existing tag information. Diverging from conventional methods, their approach harnesses the concept of webpage similarity to facilitate tag propagation. Tags are shared based on their weight and the degree of similarity between the source and recipient webpages. This similarity is quantified using a metric that integrates four cosine similarity measures, effectively encompassing both tag attributes and content. In response to the evolving Web 2.0 landscape, Song et al. [3] focus on tagging, where users assign custom keywords to digital content. They address the challenge of automating effective tag recommendations on social networks. Their study introduces two innovative document-centered methodologies: a graph-based method using bipartite graphs to uncover document topics and a prototype-based approach using a sparse multiclass Gaussian process classifier to identify representative documents. Both models employ novel tagranking mechanisms within topic clusters or classes. Empirical validation on Del.icio.us, CiteULike, and BibSonomy datasets demonstrates the superiority of document-centered models over user-centric and established topic models like LDA and SVM classifiers. Suchanek et al. [4] quantitatively evaluate two assumptions in social tagging: tag meaningfulness and the influence of tag suggestions. They analyze tag semantics by examining their alignment with content on popular web pages, finding a potential correlation with page popularity. They also studied the impact of tag suggestions on user tagging behaviors, involving over 4,000 participants, and discovered that around one-third of tag applications may be influenced by tag suggestions. These findings have important implications for improving the effectiveness of social tags in various applications, including search and information extraction. In their work, Mohideen et al. [5] introduce a novel graph-based indexing approach to enhance search latency for Boolean AND queries (BAQ). They use a hash table to represent a graph structure, reducing the need for multiple intersections

during BAQ execution. Their performance analysis compared the new method to the Inverted Index, a common structure for textual documents. The study also included a comparison with Elasticsearch, an enterprise-level search engine based on the Inverted Index. The results confirm significant latency reduction with the graph-based indexing technique for BAQ, offering promise for text retrieval optimization. El-Gayar et al. [6] highlight the growing reliance on search engines for information retrieval but note challenges in delivering precise results for ambiguous queries. They propose a comprehensive search engine framework that combines keyword-based and semantic ontology-based approaches, introducing a fuzzy membership-based ranking algorithm and a mathematical model to reveal semantic connections between keywords. Extensive testing across eight cases shows the frame-work's effectiveness, achieving a 97% precision rate with efficient response times compared to other systems. In the face of the internet's data explosion, Helin et al. [7] introduce an innovative unstructured big data platform for efficient data retrieval. In [8-16] several frameworks in support of the literature of the proposed work have been depicted.

Proposed System Architecture

Fig.1 shows the proposed system architecture for the web page tag recommendation framework encompassing meta heuristic optimization and semantic artificial intelligence. The web page dataset is subjected to extraction of the terms and categories because the dataset taken here is categorical in nature. The term frequency-inverse document frequency is also applied from the web page dataset in order to deliver the informative terms.

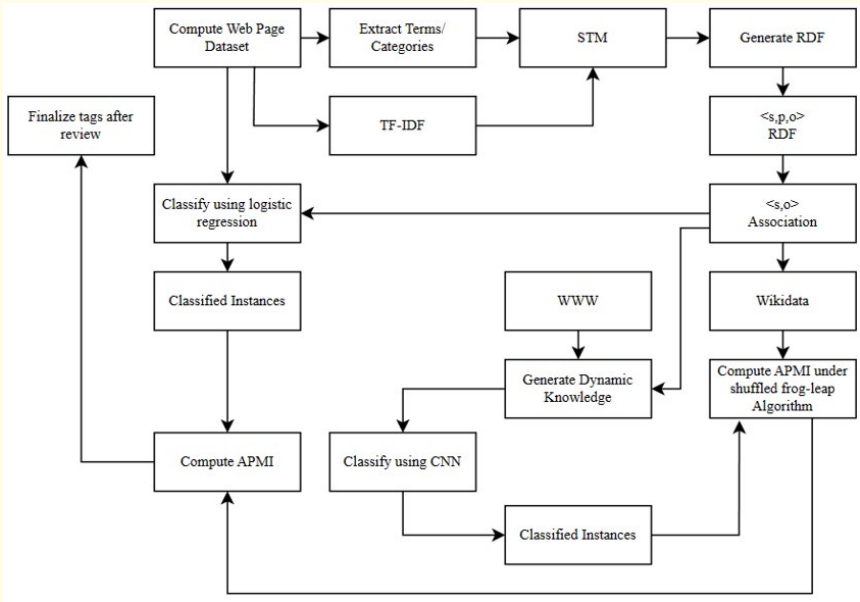

Figure 1: Proposed system architecture for Web Page Tagging.

The web page dataset is also subjected to the application of TF-IDF (term frequency-inverse document frequency) model in order to derive informative terms which are frequent within the document corpus and the terms which are rare across the document corpus. The web page dataset is considered as an individual document corpus and the terms which come out of this particular pipeline are subjected to structural topic modelling (STM) which is a topic modelling framework. This framework increases the topics laterally by using the world wide web itself as the reference corpora. The term frequency of a term or phrase refers to the number of occurrences of that specific term within a document, relative to the total number of terms present in the document. It is given by Eq. (1).

$$TF = \frac{Number\ of\ times\ the\ term\ appears\ in\ the\ document}{Total\ number\ of\ terms\ in\ the\ document} \quad (1)$$

The Inverse document frequency (IDF) depicted as Eq. (2) of a term indicates the relative occurrence of the term within the corpus. Specialized terms specific to a limited number of documents (e.g., technical terminology) are assigned greater significance compared to frequently appearing words across all documents in the corpus (e.g., a, the, and).

$$IDF = log \frac{Number\ of\ \ documents\ in\ the\ corpus}{Number\ of\ documents\ in\ the\ corpus\ containing\ the\ term} \quad (2)$$

Structural Topic Modeling (STM) is an innovative approach in natural language processing that combines topic modeling and regression analysis. It enables researchers to uncover hidden thematic patterns in text data while considering document-level metadata, making it a valuable tool for gaining insights from extensive text corpora across various fields and extracting meaningful relationships between topics and co-variates to understand textual structures better. The entities which come out of this pipeline is subjected to the generation of RDF using a tool called the Photo RDF-Gen tool. The Resource Description Framework (RDF) is generated using the Photo RDF-Gen which is a Dublin core RDF generation tool or framework. The RDF is in a triadic format with subject, predicate, object in the <s, p, o> format. The <s, p, o> format is a complicated format because the subjects and objects can be terms or sentences whereas the predicate can be a link. Due to the approachability the predicate is dropped because the predicate is only a link between the subject and the object. Therefore, the subject and object itself can yield lateral semantics.

Hence the RDF subject-object association is only considered by dropping the predicate part and this is fed as features to the Logistics Regression Classifier (LRC) to classify the dataset itself and yield the classified instances. Subsequently the RDF subject and object terms are used to crawl the dynamic knowledge stack from the world wide web via an agent which is designed using AgentSpeak.

The state of the agent is to use the subject and object terms separately or together and yield dynamic knowledge from the web crawler of the world wide web. The behavior of the agent is to assimilate knowledge from several pockets of the world wide web and create a knowledge stack. This knowledge stack is extensively large and is derived from the web and it has to be classified and this classification is achieved using the CNN classifier which is a strong and powerful deep learning classifier. This classifier works on the principle of automatic feature selection or auto handcrafted feature selection. The entities which come out of the CNN classifier are the classified instances and these classified instances are subjected to the computation of Adaptive Pointwise Mutual Information (APMI) measure with that of the entities which come out of the Wikidata API to which the subject and the object of the RDF is fed as primary inputs.

The Wikidata API gathers knowledge from the web. The entities resulting from the APMI measure computation are dynamic PMI measures with a threshold of 0.5 due to their complex nature, as it follows a two-step Pointwise Mutual Information paradigm based on adaptivity coefficient. The shuffled frog-leap algorithm optimizes intermediate steps using APMI as an objective function. It produces more optimal solution sets, which are then evaluated with the same APMI threshold as the classified instances. The threshold remains unchanged due to the strength of the APMI measure and the subjective nature of the optimal entities. The instances that come out of the APMI pipeline are filtered with an immediate threshold of 0.5 to yield final tags which is prioritized in the increasing order of the APMI measure, and this is sent for review after which the tags are finalized. The review is done by domain experts. The finalization of tags takes place on the prioritized review. This process of finalization of the tags takes place by correlating the finalized tags with that of the terms or categories in the web pages.

Convolutional Neural Networks (CNNs) have revolutionized computer vision and other domains by automatically extracting hierarchical features from images. They use convolutional layers with small filters to detect patterns and abstract features, making them powerful in image classification, object detection, facial recognition, and even natural language processing when combined with other models. Their success high-lights their role in deep learning. The PMI and Adaptive PMI measures are used to calculate semantic similarity, with the latter being a novel approach for assessing semantic heterogeneity. Eq. (3), (4), and (5) depict the PMI, APMI and the

adaptivity coefficient for APMI respectively.

$$pmi(m,n) = h(m) + h(n) - h(m,n) \quad (3)$$

$$APMI(m,n) = \frac{pmi(m,n)}{p(m)(n)} + y \quad (4)$$

$$y = \frac{1+\log[p(m,n)]}{p(n)\log[p(m)] - p(m)\log[p(n)]} \quad (5)$$

The Adaptive Pointwise Mutual Information (APMI) measure is an improved version of the PMI measure used to calculate semantic similarity. It outperforms other PMI variants, such as the Normalized PMI as depicted in Eq. (6) strategy, due to its utilization of an adaptive coefficient 'y.' This adaptive coefficient, as defined in Eq. (5), involves a logarithmic quotient in both its numerator and denominator. When combined with the PMI value, the adaptive coefficient enhances the system's overall performance. APMI boosts confidence in computing semantic heterogeneity, resulting in higher relevance for the pages returned to the user. Pointwise mutual information can be normalized to a range between -1 and +1. A value of -1 indicates that the items never occur together, 0 implies independence, and +1 represents complete co-occurrence.

$$npmi(m,n) = \frac{pmi(m,n)}{h(m/n)} \quad (6)$$

Where $h(m, n)$ is the joint self-information $-log_2 p(m, n)$.

Logistic regression is used for binary classification, estimating the probability of an observation belonging to a specific class using the sigmoid function. It's applied in fields like medicine, finance, and social sciences to predict binary events and identify contributing factors. Logistic regression offers interpretability through the analysis of input feature coefficients, revealing their direction and strength of influence on class likelihood. It accommodates both continuous and categorical features, adapting to diverse datasets. Despite its simplicity, logistic regression is effective, particularly when the relationship between predictors and the binary outcome is roughly linear. However, practitioners must ensure assumptions of independence and linearity are met for optimal performance. It remains a fundamental and interpretable method widely applied in binary classification tasks. The Shuffled Frog Leap Algorithm (SFLA) is a metaheuristic inspired by frogs' leaping behavior, designed to efficiently solve complex optimization problems. It models frog leaps as local search operations and promotes exploration and exploitation through a shuffling step. SFLA excels in numerical optimization, feature selection, and engineering design. It is efficient, adaptable to large datasets, easy to implement, and supports parallel processing. SFLA's balance between exploration and exploitation helps find high-quality solutions, making it suitable for various real-world optimization tasks.

Performance evaluation and Results

The experimentations were conducted on a single large integrated dataset comprising of four distinct constituent datasets. These datasets are namely the OFEPHI Web Page dataset, MINEM web page-SEAL dataset, PWC web page dataset and the Web-sites using Web Page Maker in Denmark dataset. These four datasets are integrated into a single large-scale dataset by using a common annotations tool to generate the annotations and pre-prioritizing the entire dataset based on the web pages. Subsequently an automized crawler also is used to crawl the web pages from the current structure of the live web 3.0 and include the URLs as well as the categories and the keywords of the webpages as annotations into the indicated single large dataset on which the experimentations were conducted.

The SOTW framework's performance is evaluated using precision, recall, accuracy, F-measure percentages, and False Discovery Rate (FDR). The SOTW model shows the highest average precision (95.18%), recall (97.45%), accuracy (96.31%), and F-measure (96.30%), with the lowest FDR value (0.05). These metrics assess result relevance, while FDR quantifies false positives. The SOTW model is benchmarked against similar models like WPPR, CBTR, and ATRS. The WPPR model yields an overall average precision percentage of 89.15%, overall average recall percentage of 91.14%, an average accuracy percentage of 90.14%, an average F-measure percentage of 90.13% and with an FDR of 0.11. Similarly, the CBTR model yields an average precision percentage of 90.13%, an average recall percentage of 93.48%, an average overall accuracy of 91.80%, an average F-measure value of 91.77% and an FDR value of 0.10. Likewise, the ATRS model yields an average precision percentage of 92.09%, an average recall percentage of 94.36%, an average accuracy of 93.22%, an average F-measure value of 93.21% and a low FDR value of 0.08.

Model	***Average Precision %***	***Average Recall %***	***Average Accuracy %***	***Average F-Measure %***	***FDR***
WPPR [1]	89.15	91.14	90.14	90.13	0.11
CBTR [2]	90.13	93.48	91.80	91.77	0.10
ATRS [3]	92.09	94.36	93.22	93.21	0.08
Proposed SOTW	95.18	97.45	96.31	96.30	0.05

Table 1: Comparison of Performance of the SOTW with other baseline models.

From Table 1 it is inferable that the proposed SOTW model yields the highest average precision percentage, highest average recall, highest average F-measure percentage, and lowest FDR value of 0.05 due to the reason that the proposed SOTW model is a semantics driven model for tag recommendation for web pages. The proposed model features a robust learning infrastructure, including lightweight logistic regression and machine learning classifiers. It incorporates structural topic modeling (STM) for topic generation, RDF for subject-predicate associations, and introduces Adaptive Pointwise Mutual Information (APMI) measures for semantics-oriented relevance computation. Convolutional neural networks classify the dynamic knowledge stack from the web. This dynamic knowledge stack, combined with RDF, STM, TF-IDF, and Wikidata entities, enriches the model's knowledge density. The shuffled frog-leap algorithm optimizes the framework for tag recommendation. Overall, the SOTW framework excels in web page tagging.

The WPPR model's performance issues stem from its lack of semantic logic regression and CNN models, affecting its expected performance, relying solely on keyword frequency and their association with meaningful interrogative words. The CBTR model underperforms due to its simplistic tag recommendation approach, overlooking advanced techniques and using limited cosine similarities. In contrast, the SOTW model offers complexity and advanced techniques, making it more effective for untagged web pages. The ATRS model focuses on enhancing tag recommendation but lacks a comprehensive comparison with established models. The SOTW framework outperforms both the CBTR and ATRS models. The precision vs the number of recommendations distribution curve is depicted in Fig.2.

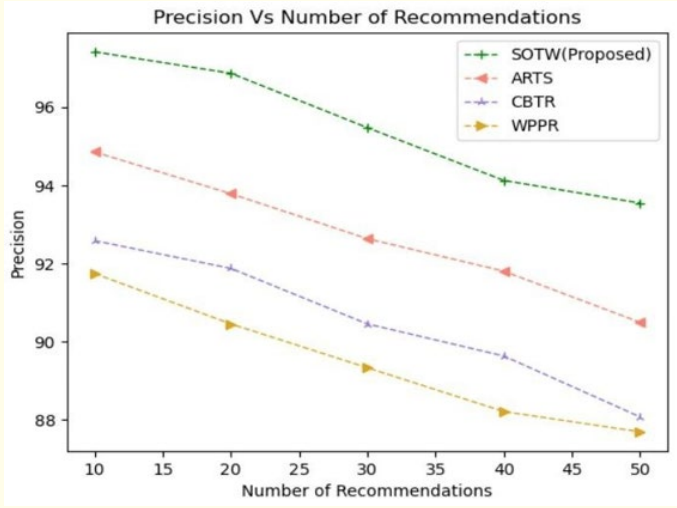

Figure 2: Precision vs Number of recommendations of the proposed SOTW and other baseline models.

Conclusion

This paper presents a knowledge-centric framework for webpage tagging, extracting informative and rare terms from webpage data. It employs structural topic modeling (STM), RDF, and the Wikidata Pipeline for knowledge augmentation, significantly enhancing precision. The RDF subject-object retention enriches lateral co-semantics. The model includes LRU and CNN deep learning classifiers for strong term classification. The APMI measure computes semantic relatedness, and the shuffled frog-leap algorithm optimizes solutions. The framework achieves an outstanding precision of 95.18%, a low False Discovery Rate (FDR) of 0.05, and an impressive F-measure of 96.30%, establishing it as the top model for semantics-oriented webpage tagging compared to baseline models.

References

1. Sharma R., et al. "Web page indexing through page ranking for effective semantic search". In 2013 7th International Conference on Intelligent Systems and Control (ISCO) (2013): 389-392.
2. Lu YT., et al. "A content-based method to enhance tag recommendation". In 21 international joint conference on artificial intelligence (2009).
3. Song Y, Zhang L and Giles CL. "Automatic tag recommendation algorithms for social recommender systems". ACM Transactions on the Web (TWEB) 5.1 (2011): 1-31.
4. Suchanek FM, Vojnovic M and Gunawardena D. "Social tags: meaning and suggestions". In Proceedings of the 17th ACM conference on Information and knowledge management (2008): 223-232.
5. Mohideen AK., et al. "A Data Indexing Technique to Improve the Search Latency of AND Queries for Large Scale Textual Documents". In 2020 IEEE/ACM International Conference on Big Data Computing, Applications and Technologies (BDCAT) (2020): 37-46.
6. El-Gayar MM., et al. "Enhanced search engine using proposed framework and ranking algorithm based on semantic relations". IEEE Access 7 (2019): 139337-139349.
7. Helin Z., et al. "High-Speed Retrieval Method for Unstructured Big Data Platform Based on K-Ary Search Tree Algorithm". In 2022 IEEE Conference on Telecommunications, Optics and Computer Science (TOCS) (2022): 598-601.

8. Houssein EH., et al. "Semantic Protocol and Resource Description Framework Query Language: A Comprehensive Review". Mathematics 10.17 (2022): 3203.
9. Ledentsov A. "Knowledge Base Reuse with Frame Representation in Artificial Intelligence Applications". IAIC Transactions on Sustainable Digital Innovation 4.2 (2023): 146-154.
10. Canitrot M., et al. "The KOMODO system: Getting recommendations on how to realize an action via Question-Answering". In Proceedings of the KRAQ11 Workshop (2011): 1-9.
11. Jagan ND, Deepak G and Santhanavijayan A. "MPTR: A Metadata-Driven Prospective Tag Recommendation Model Based on Integrative Classification". In Advances in Data and Information Sciences: Proceedings of ICDIS 2022 (2022): 291-301.
12. Deepak G and Santhanavijayan A. "OntoBestFit: a best-fit occurrence estimation strategy for RDF driven faceted semantic search". Computer Communications 160 (2020): 284-298.
13. Pushpa CN., et al. "Onto Collab: Strategic review oriented collaborative knowledge modeling using ontologies". In 2015 Seventh International Conference on Advanced Computing (ICoAC) ((2015): 1-7.
14. Deepak G, Rooban S and Santhanavijayan A. "A knowledge centric hybridized approach for crime classification incorporating deep bi-LSTM neural network". Multimedia Tools and Applications 80.18 (2021): 28061-28085.
15. Rithish H, Deepak G and Santhanavijayan A. "Automated assessment of question quality on online community forums". In International Conference on Digital Technologies and Applications (2021): 791-800.
16. Kumar N, Deepak G and Santhanavijayan A. "A novel semantic approach for intelligent response generation using emotion detection incorporating NPMI measure". Procedia Computer Science 167 (2020): 571-579.

Volume 7 Issue 3 September 2024