International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

ISSN:2147-6799 www.ijisae.org Original Research Paper

Weighted Hashing-Based Capture Text Similarity Estimation with the Cross-Media Semantic Level

Lamhot Naibaho[1], Kuldeep Singh Kaswan[2], Pankaja R[3], Shrishailappa Patil[4], Santosh Mitkari[5], Dr. Vivek Nivruttirao Waghmare[6]

Submitted: 20/08/2022 **Accepted:** 23/11/2022

Abstract: Web Mining is an emerging trend for the drastic advancement of the different data mining techniques. The web mining process comprises the sequence of operations that are comprises of the different languages those need to be processed effectively. The estimation of the similarity between the ontologies words and the sequences are computed. This paper proposed a Weighted Hashing Similarity Estimation (WHSE). The proposed WHSE model comprises of the weighted values for the estimated semantics. The computed semantics are updated in the hashing table for the estimation of the features in the variables. The proposed WHSE computes the similarity score for the extracted sematic word features in the ontology and computes the key words. The proposed WHSE model performance is comparatively examined with the existing technique. The measured recall, precision and accuracy value expressed that proposed WHSE achieves the 0.98 accuracy value for the semantic ontology. The comparative analysis expressed that proposed WHSE achieves the ~3% - 7% improvement than the existing technique for the semantic level.

Keywords: *Web Mining, Semantic Level, Weighted Hashing, Text, Similarity Value, Cross-Media*

1. Introduction

Web mining is the procedure of discovering patterns from the WWW, which is one of the applications of data mining techniques. Web mining techniques play a crucial role in dealing with the information overload problem in large-scale data collection [1]. Most notably, the Web mining technique belongs to the information retrieval process with the data mining technique to find the data patterns that are desired by the user [2]. Based on the requirement of different users, large, dynamic and unstructured new web pages are being developed continuously. Since then many technologies that include the web mining algorithm and many traditional data mining algorithms are employed to analyze a collection of large amount of data in the weblog [3]. By exploring the web pages and acquiring the required information accurately, Web mining improves the performance of the IR process. Web mining comprises of the three categories such as web usage, structure mining and content in web mining.

[1] *Dr., S.Pd., M.Hum., Lecturer/Researcher, English Language Education Study Program, Faculty of Letters and Languages, Christian University of Indonesia, Email: lamhot.naibaho@uki.ac.id, 0000-0001-9893-7165*
[2] *Professor, Department of CSE, Galgotias University, Greater Noida Uttar Pradesh, India.*
Email: kuldeep.kaswan@galgotiasuniversity.edu.in, Orcid Id-0000-0003-0876-0330
[3] *Assistant Professor, Information Science and Engineering, Sri Venkateshwara College of Engineering, Bengaluru, India.*
Email: pankaja.ssu@gmail.com, 0000-0001-5752-8023
[4] *Professor, Computer, Vishwakarma Institute of Technology, Pune, India*
Email: patil.st@vit.edu, https://orcid.org/0000-0002-9440-3446
[5] *Assistant Professor, Mathematics, Bharati Vidyapeeth's College of Engineering for Women, Pune, India.*
Email: santosh.mitkari@bharatividyapeeth.edu, 0000-0002-4689-5044
[6] *Associate Professor, Information Technology, SITRC, Nashik, India*
Email: dr.vnwaghmare@gmail.com

Due to the dramatic growth of the web content, the IR has become a critical task in the real world. Moreover, the massive amount of data on the Internet also has turned the knowledge management and information access as extremely challenging tasks [4]. The keyword-based IR techniques are traditional and also the most popular tool to retrieve the relevant results from the WWW, even though, these techniques do not cope up in delivering the related information when there is a continuous evolution of the web data [5]. The keyword-based technique has to match the user query with the data on the web and retrieve the result which has one or more query terms specified by the user [6]. As a result, it delivers irrelevant information due to the inadequate information on the keywords of the user query. Thus, the search engine needs to understand the intention of the user query and the exact context of the query terms to improve the search accuracy [7]. An additional consideration of a semantic dimension to the conventional IR techniques that assist the search tool to provide the intelligent and relevant information from the massive amount of web data [8]. The semantic level based information retrieval system identifies the relevant keywords and also determines the meaning of the query terms, which facilitates the system to retrieve all the related information regarding the user query [9]. Moreover, the semantic similarity measure creates an impact on numerous potential data mining applications such as paraphrase recognition, Word Sense Disambiguation (WSD), document retrieval, malapropism detection, and text categorization.

In day to day scenario the language keep on altered with the evolution of the new concept and trends in the time [10]. The dynamic change in the language exhibits the similarity in the certain aspects based on the estimated features. The dynamic updation of the concept and idea it is necessary to develop a effective scheme to derive significant results [11]. Through semantic analysis the linguistic resources are computed based on the different WordNet and Latent Semantic Analysis (LSA) for

the information update. The information-oriented society comprises of the socia media those are effectivelt involved in the collection of knowledge based on the emerging trends and sources to exhibits significant results for the medical resources with estimation of the similarity semantic value with computation of error [12 – 14]. This paper presented a cross language sematic estimation model for the ontology processing in the similarity estimation.

2. Related Works

In information retrieval, understanding the intent of the user need is essential to retrieve the relevant information over the huge amount of data. Hence, in order to extract relevant information determining the meaning and semantic relatedness between every word in the query is essential. The semantic similarity plays an inevitable role in measuring the relatedness between the terms. This subsection presents a survey of some of the existing systems exploits the semantic similarity measures to improve the information retrieval system. An application-oriented evaluation of the semantic relatedness measurement uses the WordNet ontology that depends on the distributional similarity of the lexical resource [14].

The sibling discovery approach extends the ontology by discovering the new terms in a sibling relationship with the existing terms of an ontology in a semi-automatic fashion. This method exploits two techniques for extracting the new terms from the web. The initial approach finds the HyperText Markup Language (HTML) document structure, and the next approach exploits the text mining to extract the siblings. Search query given by the user is inadequate to represent the need of the user as people from different background, knowledge, and expectation thus leads to reduce the performance of information retrieval system. Hence, the complement keywords which are semantically mapped with the query terms are exploited instead of the incomplete keywords in the information search [15].

In order to overcome the limitations of keyword match IR techniques, the Query expansion approach appends the semantically related terms to the query which overcomes the word mismatch problem and WHSEs the result of IR. In the query expansion, the approach which uses the lexical resources such as WordNet to select the expansion term for the query is known as lexical based approaches. Some of the approaches exploit statistical measures such as co-occurrence measures, or lexical co-occurrence measures to select the expansion terms are known as statistical approaches. In [16] discusses the Automatic Query Expansion with a large number of approaches and its benefits. Semantic similarity measures estimate the similarity between the concepts by extracting the data from the lexical knowledge sources. The WordNet is a widely used knowledge source for semantic similarity estimation. Several approaches exploit Wikipedia as a knowledge source to estimate the semantic similarity between the terms in order to improve the accuracy of information retrieval [17]. The updated knowledge source web is also playing a considerable role in measuring semantic similarity between concepts which helps to WHSE the result accuracy of IR due to the lack of concept coverage in single knowledge source, the semantic similarity approaches utilize more than one knowledge source to estimate the semantic similarity between the terms (Zhao et al., 2012).

3. Weighted Hashing Technique for the Similarity Estimation

Initially, the WHSE approach partitions the disease ontology into anatomy ontologies based on the human anatomy using the ontology modularization method, which eases the ontology updation. Also, the diseases that infect the specific human anatomy are comprised with each anatomy-ontology. The Vector Space Model (VSM) technique assists the proposed approach in discovering the candidate anatomy ontologies by matching the input term with anatomy ontology. The overall process of the proposed WHSE is presented in figure 1.

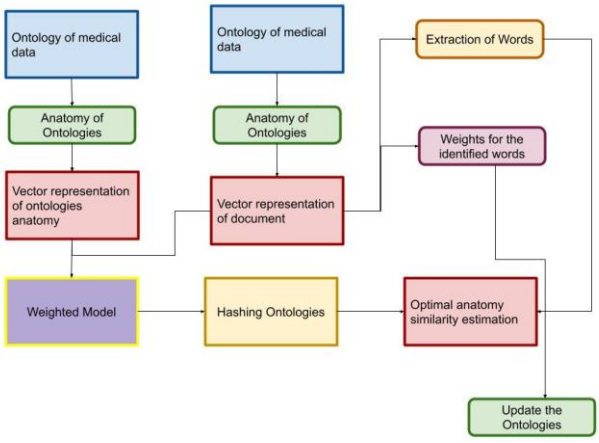

Figure 1: Architecture of the WHSE

Algorithm 1: Ontology Word-Pair features
Input: Compute the Ontology values (Hname), Disease Ontology (Dis_Onto)
Output: Anatomy-Ontologies
Procedure:
Graph = Dis_Words;
Hum_Ana[] = Comprises of the human anatomy;
For (Hum_Ana[T1, T2,...Tn]) {
Module Extraction (Hname, Dis_Onto)
{
If ((Hname = = Class of Dis_Onto) || (Hname = = SubClass of Dis_Onto) || (Hname = = Attributes of Dis_Onto))
{
Extract the label classes, sub-classes and ontology properties
}
}
Initialize Mi=0; words (w)=0; generated Word-pair (WP)=0;
While
Compute the message form(m)-> document in the chat (Mi)
do
m -> m++;
for
for i={1,...n} do
Wi(7) -> compute the words in the messages
m -> {Wi(7),....Wn(7)}
Wi(7) -> WPn
C 1 <- {C11 ,.....Cn 1 }
Endfor
//Semantic word-pair generation
for all WPn do
WPn -> WPn –
Endif
Compute the list values (a',a'',...an) -> motive WP(a)
WP <- measure the similar word-pair

end if
end for
end for
endwhile

The algorithm 1 is constructed to select the relevant modules from Disease Ontology. It employs the human anatomy entities to partition the Disease Ontology. In the beginning, the disease ontology is represented as the graph model. The WHSE approach maps the human anatomy with the entities appearing in the ontology, and its matching class, sub-class, and attributes are extracted with its relations as the anatomy ontology graph. However, when the human anatomy term "eye" is mapped to the entity of ontology entity "eye, lymphoma" by a partial mapping of the term. Hence, it has been comprising with the eye anatomy-ontology. , the WHSE system fix 0.5 value for partial mapping is 0.5 which assist in minimizing the deviation. In equation (1), each anatomy-ontology weight has been calculated. According to the weight of anatomy ontology, the ranking is done in descending order.

$$Weight\ (SV_n) = \sum_{i=1}^{m} \left(V_n(t_i)\right) \quad (1)$$

The similarity between the attributes are computed based on the estimated linkage in the semantic attributes in the ontologies anatomy as in equation (2)

$$Linkage\ Distance\ (anatomy\_ontology\ [k]) = |C_i| + \sum_{i=1}^{m}\sum_{j=1}^{n} C_{ij}\left(\frac{N_i}{H}\right) \quad (2)$$

In above equation (2) the every level term number is defined as N_i and level depth is computed as H. In the equation (2) the categories of the number is stated as $anatomy\_ontology\ [k]$ with the ontology anatomy as in n=1,2,..|n|,m, and n, represents the first level number classes for the ontology depth class, those varies between i values between 1 to m respectively as shown in figure 2.

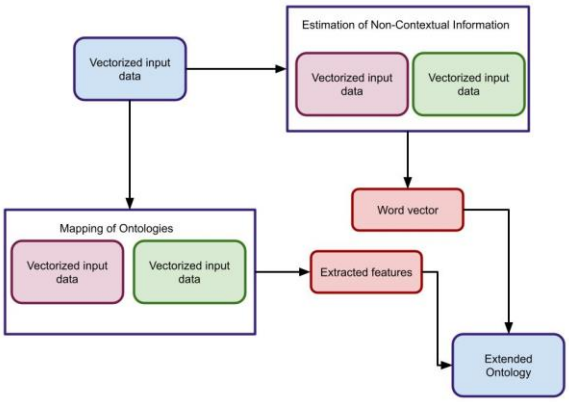

Figure 2: Architecture of the hashing model

The sematic text similarity are computed for the word-pair and any word in the estimation of the measured correlation between the words. At instances, the proposed WHSE mode compute the word-pair in the constructed n×m matrix for the different sequences. The word-pair frequencies are denoted as w_k for the time series as stated in equation (3)

$$DTW(ts1, ts2) = min[\sum_{k=1}^{K} w_k], Where\ w_k \in \{|ts1 \times |ts2||\} \quad (3)$$

The measurement of the distances are allocated based on the DTW methos through weighted hashing based similarity series. The weighted function are computed based on the covariance in the temporal features in the time series sequences $ts1$ and $ts2$ computed as in equation (4) – (6)

$$Temp_{cov(ts1,ts2)} = Cov\ (ts1, ts2) + W(f).[(ts1(I) - E(ts1)) \times (ts2(DI) - E(ts2)] \quad (4)$$

$$Corr(ts1, ts2) = \frac{COV(ts1,ts2)}{\sigma_{ts1}\sigma_{ts2}} \quad (5)$$

$$= \frac{E[(ts1-E(ts1))(ts2-E(ts2))]}{\sigma_{ts1}\sigma_{ts2}} \quad (6)$$

The covariance between the variables are estimated with the equation (6) the weighetd function for the index product is defined as $W(f)$, (I) represents the function of the index products and (DI) denoted delay index in the time series sequences. With the weighted hashing the delay index estimation is computed with the frequency difference minimum and maximal values in the time series sequences.

Dataset

The text samples are collected from the social network chat history between the time period of 2011 – 2019. The collected dataset comprises of the different phrases and hyperlinks for the blog spot, time stamp, hyperlinks and phases. The collected data is denoted as P for the URL document, post time is represented as T, the text document is phases denoted as the Q and document hyperlink denoted as L. The experimental framework employs the new disease name related document as the dataset which is extracted from the rare diseases data . Moreover, it exploits the disease ontology and the human anatomy names.

4. Experimental Setup

To implement the proposed architecture, the experimental model requires the software and hardware settings. The experimental analysis is simulated using the Ubuntu 12.04 deployed in the Inter Pentium E2160 processor with the CPU frequency of 1.80GHZ. The software requirements include Java version 1.6, Java HotSpot (TM) 64-Bit Server VM and Protege Editor.

To illustrate the performance improvement of the WHSE approach, the experimental framework compares the proposed approach with the existing approach using various scenario and performance metrics. The simulation parameters for the proposed WHSE is presented in table 1.

Table 1: Comparison of Parameters

Query Level	Precision			Recall			Accuracy		
	[10]	[11]	WHSE	[10]	[11]	WHSE	[10]	[11]	WHSE
0.1	0.73	0.82	0.93	0.66	0.73	0.89	0.67	0.84	0.87
0.2	0.77	0.82	0.94	0.67	0.71	0.92	0.67	0.86	0.93
0.3	0.83	0.84	0.93	0.68	0.74	0.93	0.69	0.88	0.94
0.4	0.85	0.86	0.92	0.71	0.73	0.95	0.72	0.87	0.95
0.5	0.87	0.85	0.93	0.73	0.76	0.96	0.73	0.85	0.97
0.6	0.86	0.86	0.94	0.72	0.79	0.98	0.74	0.84	0.96
0.7	0.83	0.87	0.95	0.68	0.81	0.97	0.76	0.85	0.97
0.8	0.85	0.88	0.96	0.69	0.83	0.96	0.77	0.89	0.96
0.9	0.89	0.87	0.97	0.73	0.82	0.97	0.81	0.92	0.95
1.0	0.87	0.86	0.98	0.75	0.80	0.98	0.83	0.93	0.96

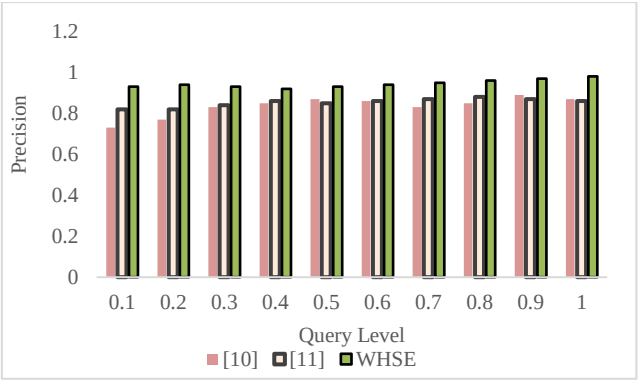

Figure 3: Comparison of Precision

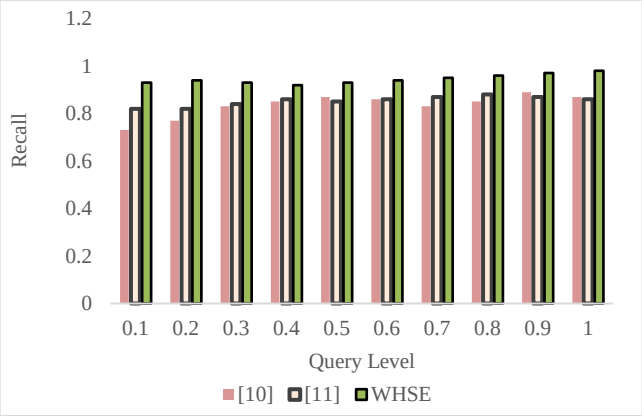

Figure 4: Comparison fo Recall

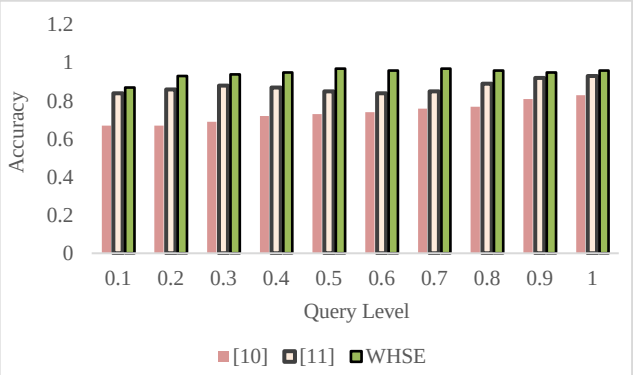

Figure 5: Comparison of Accuracy

The precision and recall of both the WHSE and the existing sibling discovery approaches are illustrated in Figure 3 and 4 respectively. The figures above show the performance with the variation of query level from 0.1 to 0.9 and the new and old terms. The query level refers that the information availability is illuminating the easy understanding of the query context. When the query level is 0.5, both the WHSE and the existing approaches obtain better precision and recall value for the old term. The WHSE approach performs better for the new medical terms than the Sibling discovery method. In the Disease Ontology, the existing approach focuses on finding the siblings in the existing terms. Hence, it leaves the new disease term. For instance, the old term and the query level is 0.9 and, the WHSE approach attains precision value by 0.98 but, the existing approach obtains only 0.97 precision value. In the same case, for the new term, the WHSE and the sibling discovery achieves 0.97 and 0.87 precision value since the WHSE approach dynamically updating the disease ontology by semantically considering the new disease terms and its symptoms in the appropriate location of the anatomy ontology.

The word pair estimated for the semantic ontology is presented in table 2 as follows

Table 2: Estimation of Word-Pair

Semantic Measure	Word pair 1	Word pair 2	Word Pair 3
1.0	0.94	0.95	0.98
1.5	0.92	0.93	0.96
2.0	0.91	0.91	0.94
2.5	0.90	0.89	0.93
3	0.89	0.88	0.92
3.5	0.88	0.87	0.91
4.0	0.87	0.86	0.9

Therefore, it is proved that the proposed approach can provide a better semantic relationship for the medical terms than other existing resource based on semantic similarity measures.

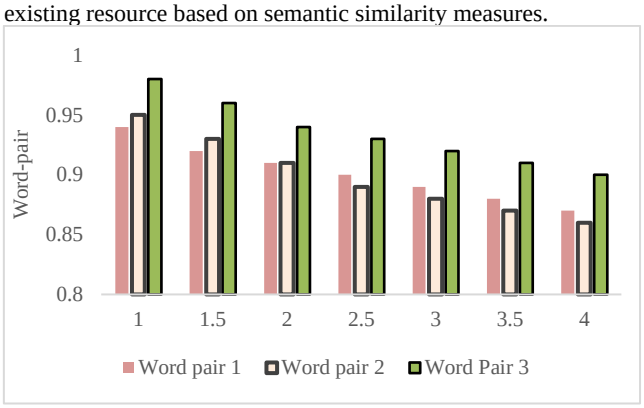

Figure 6: Comparison of Word-Pair

Figure 46 shows the performance is evaluated in terms of updating time while increasing the number of the input document and ngram value. WHSE approach exploits variation of the number of input document terms from 100 to 500 and n-gram from 1 to 4. For instance, the WHSE system obtains 0.25ms updating time when the number of terms in the document is 300 and 1-gram model. In the same case, it spends 0.38ms for the term updation when the 4-gram model. Even though the 4-gram model takes more time to update the new disease term in the disease ontology, it assists in improving the medical information retrieval by determining the multi-word new medical terms as in table 3

Table 3: Word Pairs used in Experimentation

Word Pair 1	Respiratory Disease - the Middle East Respiratory Syndrome
Word Pair 2	The respiratory syndrome in middle east – Severe acute
Word Pair 3	Yunis Varon Syndrome – The disorder in multi system

The impact of top-k anatomy ontologies selection through the F-measure. The performance of the WHSE approach escalates while varying the top-k anatomy ontologies from 3 to 9 and varying the query levels from 0.3 to 0.9. The WHSE approach increases the F-measure value when increasing the query level from 0.3 to 0.9. The inaccurate selection of optimal anatomy-ontology tends to reduce both the ontology update accuracy and semantic similarity measure accuracy. Moreover, in the determination of optimal anatomy ontology, the number of anatomy ontologies involved is the primary factor as the number of top-k ontologies varies the f-measure. While the top-2 or 3 ontologies are involved in optimal anatomyontology selection, there is a possibility to neglect the more related anatomy ontology. The selection of top-k anatomy ontologies is of paramount importance for the new disease updation in the disease ontology to improve the medical information retrieval. As a result, the WHSE approach obtains the 0.913 F-measure value when there are top-3 anatomy ontologies, and query level is 0.9

5. Conclusion

Web mining based approaches are estimated based on the evaluated features those need to be processed and evaluated for the similar languages. Web mining comprises of the different language features for the computation of the variables. The proposed WHSE computes the weighetd vaues for the estimation of the features. The

weighted features are updated in the hashing tables for the extracted word-pair features in the languages. The proposed model exhibits the higher precision, recall and accuracy value of 0.98 which is significantly minimal than the existing techniques.

References

[1] Kulmanov, M., Smaili, F. Z., Gao, X., & Hoehndorf, R. (2021). Semantic similarity and machine learning with ontologies. Briefings in bioinformatics, 22(4), bbaa199.

[2] Chandrasekaran, D., & Mago, V. (2021). Evolution of semantic similarity—a survey. ACM Computing Surveys (CSUR), 54(2), 1-37.

[3] Ishiguro, S., & Saito, S. (2021). The detrimental effect of semantic similarity in short-term memory tasks: A meta-regression approach. Psychonomic Bulletin & Review, 28(2), 384-408.

[4] Zad, S., Heidari, M., Hajibabaee, P., & Malekzadeh, M. (2021, October). A survey of deep learning methods on semantic similarity and sentence modeling. In 2021 IEEE 12th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON) (pp. 0466-0472). IEEE.

[5] Hayes, T. R., & Henderson, J. M. (2021). Looking for semantic similarity: what a vector-space model of semantics can tell us about attention in real-world scenes. Psychological Science, 32(8), 1262-1270.

[6] Lu, K., Li, R., Chen, X., Zhao, Z., & Zhang, H. (2021). Reinforcement learning-powered semantic communication via semantic similarity. arXiv preprint arXiv:2108.12121.

[7] Mohammed, S. M., Jacksi, K., & Zeebaree, S. (2021). A state-of-the-art survey on semantic similarity for document clustering using GloVe and density-based algorithms. Indonesian Journal of Electrical Engineering and Computer Science, 22(1), 552-562.

[8] Zhang, P., Huang, X., Wang, Y., Jiang, C., He, S., & Wang, H. (2021). Semantic similarity computing model based on multi model fine-grained nonlinear fusion. IEEE Access, 9, 8433-8443.

[9] Jin, L., & Gildea, D. (2022, May). Rewarding Semantic Similarity under Optimized Alignments for AMR-to-Text Generation. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) (pp. 710-715).

[10] Kocoń, J., & Maziarz, M. (2021). Mapping WordNet onto human brain connectome in emotion processing and semantic similarity recognition. Information Processing & Management, 58(3), 102530.

[11] Li, S., Abel, M. H., & Negre, E. (2021, May). Ontology-based semantic similarity in generating context-aware collaborator recommendations. In 2021 IEEE 24th International Conference on Computer Supported Cooperative Work in Design (CSCWD) (pp. 751-756). IEEE.

[12] Yang, F., Liu, Y., Ding, X., Ma, F., & Cao, J. (2022). Asymmetric Cross-Modal Hashing with High-Level Semantic Similarity. Pattern Recognition, 108823.

[13] Vetter, D., Tithi, J. J., Westerlund, M., Zicari, R. V., & Roig, G. (2022). Using Sentence Embeddings and Semantic Similarity for Seeking Consensus when Assessing Trustworthy AI. arXiv preprint arXiv:2208.04608.

[14] Netisopakul, P., Wohlgenannt, G., Pulich, A., & Hlaing, Z. Z. (2021). Improving the state-of-the-art in Thai semantic similarity using distributional semantics and ontological information. Plos one, 16(2), e0246751.

[15] Solomon, S., Cohn, A., Rosenblum, H., Hershkovitz, C., & Yamshchikov, I. P. (2021). Rethinking Crowd Sourcing for Semantic Similarity. arXiv preprint arXiv:2109.11969.

[16] Kakad, S., & Dhage, S. (2021, March). Ontology construction from cross domain customer reviews using expectation maximization and semantic similarity. In 2021 International Conference on Emerging Smart Computing and Informatics (ESCI) (pp. 19-23). IEEE.

[17] Wang, L., Zhang, F., Du, Z., Chen, Y., Zhang, C., & Liu, R. (2021). A hybrid semantic similarity measurement for geospatial entities. Microprocessors and Microsystems, 80, 103526.