EAI Endorsed Transactions
on Energy Web

Research Article EAI.EU

Novel Semantic Relatedness Computation for Multi-Domain Unstructured Data

Rafeeq Ahmed[1], Pradeep Kumar Singh[2], Tanvir Ahmad[1,*]

[1]Computer Engineering Department, Jamia Millia Islamia, New Delhi, India
[2]CSE Department, KNIT Sultanpur, UP, India

Abstract

Semantic Relatedness computation has been a fundamental as well as an essential step for domains like Information Retrieval, Natural Language Processing, Semantic Web, etc. Many techniques for Semantic Relatedness calculation in a single domain have been proposed. However, these techniques give inappropriate results for the massive multidomain dataset because they provide a relation between concepts across different domains, which are not related to each other. Their similarities should be minimized. In this paper, a novel method, "modified Balanced Mutual Information(MBMI)," to calculate the semantic relatedness of multidomain data has been proposed. In this proposed method, to get semantic relatedness, concepts are extracted, followed by a fuzzy vector from a given corpus. A comparison of the proposed method with other existing methods has been performed. We used medical and computer science articles as our dataset. The proposed method shows better results for multidomain data.

Received on 18 March 2020; accepted on 26 June 2020; published on 30 June 2020

Keywords: Text Mining , Semantic Similarity, Concept Extraction

doi:10.4108/eai.13-7-2018.165503

1. Introduction

Data exists mainly in three forms: structured data, semi-structured data, and unstructured data. Unstructured data comprises emails, blogs, web textual data, tweets, news articles, e-learning articles, online study material, Wikipedia, and so on. This amounts to a higher percentage than any other format of data in all open data worldwide. Because of the unstructured size, it possesses lots of ambiguity, which gives rise to different algorithms to extract information using different parameters depending on fields like news articles, web mining, spam detection, reviews, and ratings, E-learning tools do text mining for knowledge representation or information extraction. It has created tremendous revenue for areas like sentiment analysis, text summarization, movies/product recommendation systems. The source of these structured data all forms of data are Social Media, Wikipedia, News, Customer Reviews on Movies, Online Products, Foods, etc. These sources are generating Big Data, which has been well defined in [1].

Big Data is the succeeding contemporaries of business analytics as well as data warehousing and is supposed to produce top-line profits to industries. The most significant role of this marvel is the accelerated step of shift and transformation; today, as of now, it is not where we will be in merely two years. Over the last few years, data has expanded to become “unstructured” more unlike that of structured data since the corporate sector has 80 percent of data is unstructured. Also, the sources of data have engendered exceeding operational applications. Text, news, blogs, emails, e-books, geospatial, and Internet data are unstructured data. Semi-structured data is frequently an aggregate of mixed types of data that has a remarkable pattern or edifice, which is not defined as structured data.

Semantic Relatedness computation has been a fundamental as well as an essential step for domains like information retrieval, Natural Language Processing, Semantic Web, etc. Many techniques for Semantic Relatedness or Mutual Information calculations have been

*Corresponding author. Email: rafeeq.amu@gmail.com

developed like Normalized Google Distance (NGD), Balanced Mutual Information (BMI), and so on. The similarity measurement techniques should give similarity between related terms, and also it should give minimum similarity between highly dissimilar items. However, these techniques fail when we have multi-domain big data [2] because they provide a relation between concepts across the different domains, which are not related, and their similarity should be minimum. We have used medical science and computer science articles for our experiments. After preprocessing, essential concepts have been extracted for which a vector has been generated. Then fuzzy vector is obtained by using different semantic relatedness techniques.

1.1. Motivation

The motivation behind the new formula's derivation is that the existing Balance Mutual Information (BMI) computes similarity when two terms appear together and not appear together minus when one word appears, and the other is absent. However, when considering large data silos, two less correlated words appear in one part of the section, and both are missing in the significant area of data distribution. Thus BMI will give higher value, although both are less related, and this BMI will always give more value even if two terms are very less related.

1.2. Contribution

We have derived a new formula for semantic relatedness computation. We have worked on academic articles, mainly research, keeping articles from multiple domains. We have compared with existing techniques and shown our formula gives the optimized result.

1.3. Organization

Section 2 discusses semantic relatedness using different techniques. In Section 3, we have covered the proposed framework and methodology. In Section 4, the results obtained have been shown. In section 5, the Conclusion, as well as future work, is given.

2. Related Work

Computation of semantic relatedness between two concepts determines the extent to which terms are closer to each other semantically. There may exist multiple relations between two concepts, and we can also extract any association between these concepts. For example, cancer is related to chemotherapy. Relatedness is a broad term with similarity as its subset. One example is Is-A relation like cat Is-A mammal. Similarly, measures are an essential and fundamental step for information extraction/retrieval, knowledge extraction, and so on. Getting the similarities among tokens is the foremost step. The probabilistic distribution of words is used for getting mutual information semantically or lexically. Also, this can be done for sentences and paragraphs. Information retrieval is a vital field based on this concept. Many techniques have been developed to get more accurate data in Information retrieval. Latent semantic analysis is the mathematical models developed and used to improve the accuracy of information retrieval is Latent semantic indexing, also called[3]. LSA extracts a matrix of concept to concept or term-documents from a given corpus then uses Singular Value Decomposition(SVD) [4] if the matrix becomes large. The Singular Value Decomposition removes less important features this reducing the dimension of a large matrix. SVD decomposes the given S matrix of order M x N into three matrices, which in turn reduces the rank of the given matrix, thus reducing the size of a matrix as well as approximating the same information as that stored in the original matrix. Some of the semantic similarity measures are semantic indexing[5],word sense disambiguation [6, 7], or coreference resolution [8], information extraction patterns[9],topic coherence[10] , spelling correction[11]. In E-learning, Learning objects with semantic similarities are used to generate a knowledge graph and recommend a personalized learning path[12, 13].

An N-gram is applied to a set of words in a sentence in which we can say words of n-tuple with the condition that they follow each other. If we take an example of a sentence like "Hadoop is a big data tool," "Hadoop can process big data," or "Hadoop can process big data in real-time." The pattern of words following each other can be used to store it as an index.[14, 15], Damerau-Levenshtein [16, 17]. The Smith-Waterman algorithm uses protein sequence or nucleic acid sequence to find out the similar regions of the string by optimizing similarity measures by comparing segments of different lengths[18]. Needleman-Wunsch algorithm[19] also uses a string score with dynamic programming for the alignment of all possible nucleotide/protein sequences in Bioinformatics to get alignments having with the highest score. Probabilistic linkage technology has been used to link sizeable public health databases by getting the scores between two given files of individual data under uncertain environment based on probability error[20]. Given two string [21] proposed a string comparator measurement for partial computing agreement between the given strings for updating exact agreement weights if the given lines do not agree with character by character. CLEEK links entities using multidomain data, which is Chinese long-text corpus[22]. Balanced Mutual Information (BMI) calculates mutual information between two terms or concepts by taking care that how much both terms are

present or absent together minus if one term is present and second is absent and so on [23].

$$BMI(C_i, C_j) = \beta \times [Pr(c_i, c_j) \times log_2 \left\{ \frac{Pr(c_i, c_j) + 1}{Pr(c_i)Pr(c_j)} \right\} + Pr(\sim c_i, \sim c_j) \times log_2 \left\{ \frac{Pr(\sim c_i, \sim c_j) + 1}{Pr(c_i)Pr(c_j)} \right\}] + (1 - \beta) \times [Pr(\sim c_i, c_j) \times log_2 \left\{ \frac{Pr(\sim c_i, c_j) + 1}{Pr(c_i)Pr(c_j)} \right\} + Pr(c_i, \sim c_j) \times log_2 \left\{ \frac{Pr(c_i, \sim c_j) + 1}{Pr(c_i)Pr(c_j)} \right\}] \quad (1)$$

When we have a semantic context vector for two concepts, C_1 and C_2, we can find which members are both joint and distinct. The Jaccard similarity index[24] uses this method.

$$Jaccard \quad Coefficient = \frac{c_1 \wedge c_2}{c_1 \vee c_2} \quad (2)$$

Cosine similarity finds the angle between two objects by taking their features vector as input. It gives an output from 0 (not similar at all) to 1(highly similar). Another vital algorithm is Normalized Google Distance (NGD) [25], Kulback Leibler, Expected Cross-Entropy (ECH), which calculates the semantic similarity between two words as given below:

$$NGD(C_i, C_j) = \frac{max\{log_2(w_{c_i}), log_2(w_{c_j})\} - log_2(w_{c_i,c_j})}{log_2(w + 1) - min\{log_2(w_{c_i}), log_2(w_{c_j})\}} \quad (3)$$

$$KL(C_i, C_j) = \sum Pr(c_i|c_j) \times log_2 \left\{ \frac{Pr(c_i|c_j)}{Pr(c_i)} \right\} \quad (4)$$

$$ECH(C_i, C_j) = Pr(c_i) \sum Pr(c_i|c_j) \times log_2 \left\{ \frac{Pr(c_i|c_j)}{Pr(c_i)} \right\} \quad (5)$$

[26] has used the input as a short text for finding semantic similarity based on lexical matching. WordNet, being a lexical database for English, provides relations and hierarchy among synsets[27]. Other techniques using information content are Jiang and Conrath[28], Resnik [29], and Lin [30]. Wikipedia has been a vast, rapidly evolving tapestry of highly hyperlinked textual content. Wikipedia constitutes articles, categories, and redirects mostly great resource for natural language processing. Based on Wikipedia, the work has been done using Wikipedia link structure, WikiWalks[31], or Wikipedia Link Vector Model[32] and Wikirelate [33]. Semantic interpretation of terms has always been made using its vector, like word2vec or other techniques obtained through the windowing process.[34, 35] has used the Fuzzy Context vector to represent a concept. For big data analytics, a distributed technique is used in [36]. Semantic-based document clustering has been done by using Wikipedia and the concept of ontology[37].

3. Proposed Technique

3.1. Problem Statement

The prominent approaches like BMI, CP, ECH, Jaccard, KL, MI, and NGD have been used to get the semantic similarity in a given corpus for a particular domain. However, nowadays, all data sources generate Big Data, and the characteristics of Big Data have already been discussed in the Introduction part. Big Data has got massive size having data from a different domain. However, the performance of all semantic relatedness computation is not excellent as they give some similarity among terms across different domains.

3.2. Modified Balance Mutual Information – A Novel Technique

We have proposed a unique formula Modified Balance Mutual Information (MBMI) in eq five, which gives values when two terms appear together multiplied by a β factor minus (1-β) time when either two terms do not appear together. While existing Balance Mutual Information BMI computes similarity when two terms appear together and not appear together minus when either one terms appear and other is absent. However, when considering large data silos, two less correlated terms appear in one part of the section, and both are absent in significant data distribution areas. Thus BMI will give higher value, although both are less related, and this BMI will always give more value even if two terms are very less related. We have shown experimentally that our method gives optimized results.

$$MBMI(C_i, C_j) = \alpha \times [Pr(c_i, c_j) \times log_2 \left\{ \frac{Pr(c_i, c_j) + 1}{Pr(c_i)Pr(c_j)} \right\} + \delta \times [Pr(\sim c_i, c_j) \times log_2 \left\{ \frac{Pr(\sim c_i, c_j) + 1}{Pr(c_i)Pr(c_j)} \right\} + Pr(c_i, \sim c_j) \times log_2 \left\{ \frac{Pr(c_i, \sim c_j) + 1}{Pr(c_i)Pr(c_j)} \right\}] \quad (6)$$

α and δ are kept in between 0 and 1. We have kept α=0.55 and δ=0.45 to provide higher weightage if two concepts appear together in a window against if only one concept is present in a window.

4. Implementation

As we have taken research articles of the medical domain and computer domain as our input, it needs to be preprocessed to be ready for applying algorithms.

4.1. Document Preprocessing

All words with no quality of information except only grammatical connotations are involved in the removal set of words. After removing these words, the remaining set of words leaves a more productive bag of words for analysis. *POS Tagging:* The mechanism of addressing words based on their different parts of speech. *Stemming:* Various grammatical variants of a word such as noun, adjective, and adverb, the root word are called stemming. POS and different word occurrences are not considered in the formation of stemming.

$$Concept \quad C_i = \{t_1, \quad t_2 \quad \dots \quad t_m\} \tag{7}$$

Mutual information between two enitites i.e. two terms can be computed by following formula:

$$MI(t_i, t_j) = log_2 \left\{ \frac{Pr(t_i, t_j)}{Pr(t_i) * Pr(t_j)} \right\} \tag{8}$$

where proability of getting a term in a window w of term t_i's and t_j's are Pr(t_i) and Pr(t_j) respectively, estimated using w_{t_i}/w, and w_{t_j}/w where w_{t_i} and w_{t_j} are the counts of terms t_i and t_i in total windows[11].Changing the above equation from programming point of view we have We have defined a Fuzzy context vector for every concept extracted from the corpus.Thus if ith concept , lets say C_i, then membership function $\mu_(c_i)$ of t_i with C_i can be defined as

$$Concept \quad C_i = \left\{ \mu_{c_i}(t_1), \quad \mu_{c_i}(t_2) \quad \dots \quad \mu_{c_i}(t_m) \right\} \tag{9}$$

The concept C_i is having m t_i terms obtained by windowing process and corresponding weight is $\mu_(c_i)$

4.2. Semantic Computation

Finally for getting the semantic relatedness between two concepts, we require a fuzzy context vector. We have carried out an extensive computational process to get these Fuzzy vectors.

$$C_1 = \left\{ t_1(\mu_{c_1}), \quad t_2(\mu_{c_1}) \quad \dots \quad t_m(\mu_{c_1}) \right\} \tag{10}$$

$$C_2 = \left\{ t_1(\mu_{c_2}), \quad t_2(\mu_{c_2}) \quad \dots \quad t_m(\mu_{c_2}) \right\} \tag{11}$$

$$R(t_{i,c_1}, t_{match(i),c_2}) = \max_{1 \le j \le n} \{MI(t_{i_{c_1}}, t_{j_{c_2}})\} \tag{12}$$

here match(i) has been for reducing the computation process since the result will not be affected. The term in second vector with highest matching with the term in first vector will be taken for computation. Now semantic similarity between two terms can be computed as follows:

Algorithm 1 Semantic Relatedness Computation

1: **Input:** N Multi Domain Academic Articles: MDAA
2: **Output:** Fuzzy Concept Relation Matrix
3: **Concept Extraction:**
4: **Repeat**
5: Select each document d ϵ MDAA
6: Remove stop words
7: Extract the Concept C_i
8: Apply porter stemming on C_i
9: Get the frequency F_i for C_i
10: If Frequency F_i > Threshold value, Next step else skip next
11: ArrayConcept = ArrayConcept U C_i
12: **Until** All documents ϵ MDAA scanned
13: **Fuzzy Vector Generation:**
14: **Repeat**
15: Select each concept C_i ϵ ArrayConcept
16: **Repeat**
17: Select each document d ϵ DP
18: Construct text window w ϵ d
19: Calculate the joint frequency of the term T_i with C_i
20: Calculate the Fuzzy membership value using BMI, NGD, MI, NGD, MBMI, CP, ECH
21: Construct the Fuzzy Vector, F_v = {$\mu_{c_i}(t_1)$, $\mu_{c_i}(t_2)$... $\mu_{c_i}(t_m)$}
22: **Until** All documents ϵ MDAA read
23: **Until** All Concepts ϵ ArrayConcept scanned
24: **Fuzzy Matrix Generation:**
25: **Repeat**
26: Select each concept C_i ϵ ArrayConcept
27: **Repeat**
28: Select each concept C_j ϵ ArrayConcept
29: **Repeat**
30: Select each term T_i ϵ F_{v_i}
31: Select the maximum matching R_{ik} of T_i with T_k ϵ F_{v_j}, for k=0 to n.
32: Sum=$\sum R_{ik} * \mu_{c_1}(t_1) * \mu_{c_2}(t_{m(i)})$
33: **Until** All term T_i ϵ F_{v_i} scanned
34: Display the semantic distance of C_i with C_j as 1/sum
35: **Until** All Concepts ϵ ArrayConcept scanned
36: **Until** All Concepts ϵ ArrayConcept scanned

$$Similarity(c_1, c_2) = \frac{1}{m} \sum_{i=1}^{n} R(t_i, t_{m(i)}) \times \mu_{c_1}(t_i) \times \mu_{c_2}(t_{m(i)}) \tag{13}$$

where R(t_i,t_m(i)) means mutual information of t_i and $t_{m(i)}$.

$$Dist(c_1, c_2) = \frac{m}{\sum_{i=1}^{n} R(t_i, t_{m(i)}) \times \mu_{c_1}(t_i) \times \mu_{c_2}(t_{m(i)})} \tag{14}$$

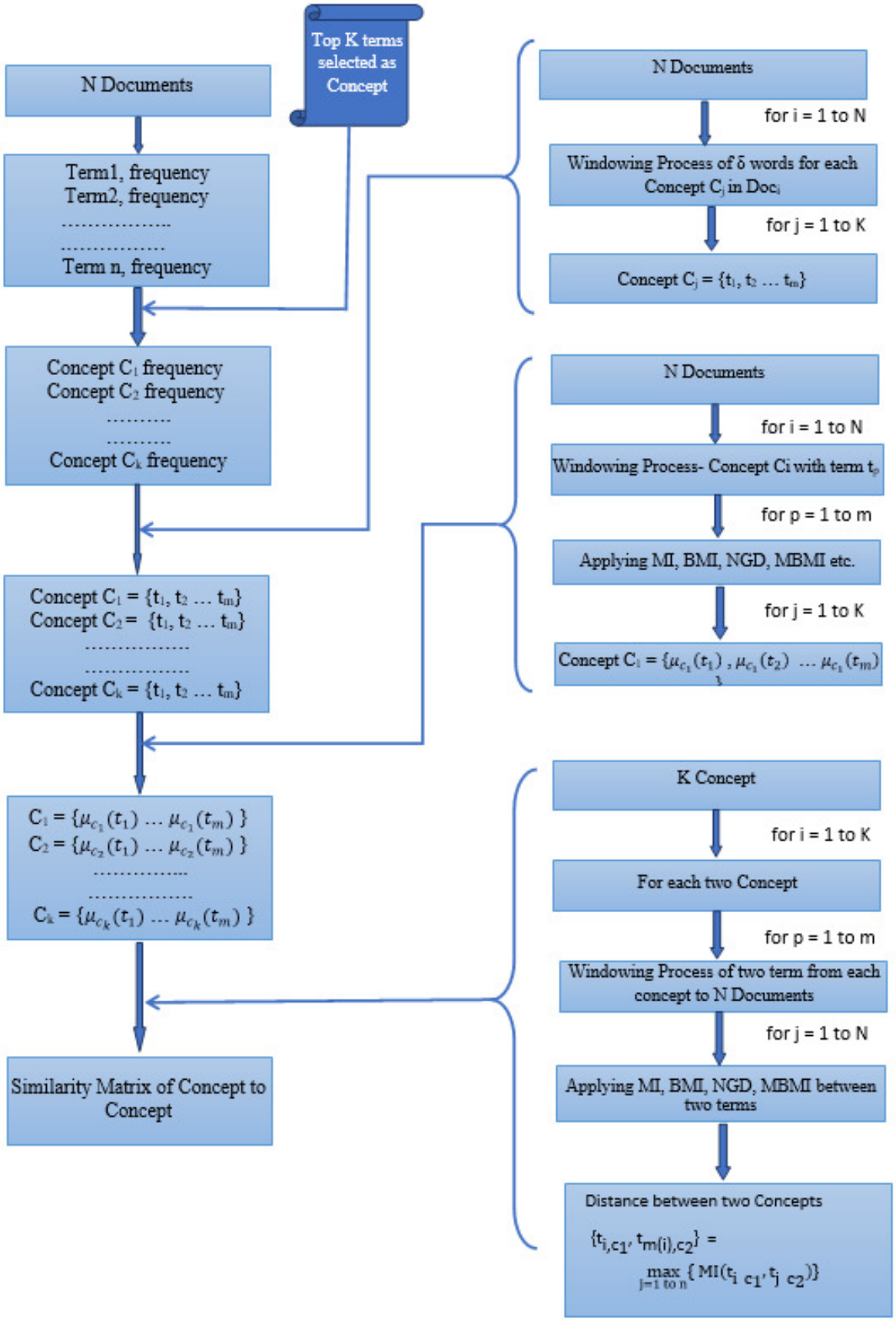

Figure 1. Flow Diagram for Semantic Similarity Matrix

IMAGE, analysis(0.642), mri(0.456), classify(0.464), registr(0.051), assist(0.225), cnn(0.002), detect(0.028), tumor(1), feature(0), application(0.195),

LEARN, machine(0.642), analysis(0.456), feature(0.464), classify(0.051), task(0.225), transf(0.002), application(0.028), detect(1), apply(0), mri(0.195

DEEP, analysis(0.774), architecture(0.522), mri(0.734), feature(0.393), apply(0.589), detect(0.566), proceeding(1), application(0.346), machine(0.367

NETWORK, architecture(0.255), proceeding(0.4), cnn(0.306), lay(1), feature(0.741), apply(0.825), arxiv(0.259), input(0), layer(0.83), niftynet(0.782),

DATA, set(0.774), augment(0.522), analysis(0.734), feature(0.393), mri(0.589), input(0.566), label(1), process(0.346), cnn(0.367), sample(0),

NEURAL, process(0.605), inform(0.623), system(0.871), artificial(0.76), architecture(0.572), predict(1), proc(0.731), recurrent(0), recogn(0.237), dete

SEGMENT, tumor(0.563), mri(0.637), automatic(0.573), cnn(0.766), ultrasound(0.294), brats(0.637), proceeding(0.618), result(1), detect(0.55), task(

MEDICAL, analysis(0.563), assist(0.637), intervent(0.573), application(0.766), machine(0.294), apply(0.637), task(0.618), proceeding(1), registr(0.55

TRAIN, dataset(0.516), set(0.083), test(0.404), pre(0.976), feature(0.869), cnn(0.628), sample(0), performance(0.295), layer(1), limit(0.759),

MODEL, machine(0), performance(0.313), task(0.313), set(0.582), architecture(0.313), cnn(0.313), analysis(0.259), feature(0.269), weight(0.586), zo

IMAG, tran(0.458), mri(0.248), modality(0.303), machine(0), phy(0.271), chen(0.326), liu(0.361), magnetic(1), resonance(0.008), analysis(0.935),

METHOD, tumor(0.215), automatic(0.236), apply(0.847), result(0.848), art(0), classify(0.774), feature(1), machine(0.872), analysis(0.147), cnn(0.154

COMPUT, vis(0.642), assist(0.456), intervent(0.464), pattern(0.051), recogn(0.225), program(0.002), aid(0.028), piscataway(1), spring(0), berlin(0.19

BAS, tumor(0), cnn(0.565), classify(0.878), predict(1), retrieval(0.334), automatic(0.776), detect(0.377), feature(0.088), mri(0.233), ultrasound(0.662),

IEEE, tran(0.718), piscataway(0.504), pattern(0.809), proceeding(0.323), vis(0.818), international(0.729), conference(0), recogn(1), biomedical(0.818

BRAIN, tumor(0.467), mri(0.888), automatic(0.849), brats(0), cnn(0.413), tissue(0.204), analysis(1), contrast(0.726), classify(0.636), multimodal(0.82

CONVOLUTIONAL, layer(0.605), lay(0.623), cnn(0.871), proceeding(0.76), detect(0.572), architecture(1), conference(0.731), international(0), net(0.23

Figure 2. Generated Fuzzy Context Vector

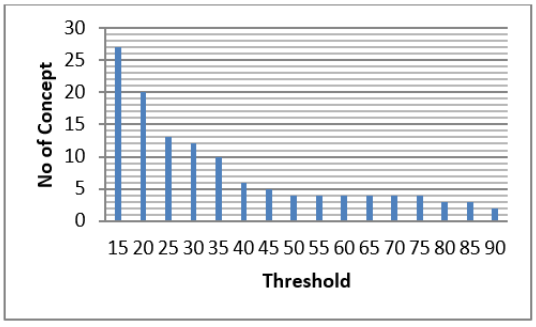

Figure 3. No of Concept extracted Vs Threshold (percentile)

5. Results And Analysis

We have taken following medical articles as our data set [38], [39],[40],[41],[43],[42] for our experiment. These articles covers deep learning, biomedicine, medical imaging, tumor image segmentation, ultrasound analysis.

These data set after being preprocessed, frequent terms have been assumed to be a learning topic in which we treat them as a concept. After applying a specific threshold value, essential concepts have been extracted. Figure 3 shows no of the concepts extracted vs. threshold value. We have set threshold values to 20 percentiles, i.e., terms with above this value have been treated as a concept for which a vector has to be generated. In this paper, a novel semantic relatedness technique has been proposed and experimented with computing semantic relatedness for cross-domain academic articles from medical and computer science journals as our data set. The result is domain-dependent, as well as the content of articles being taken as input. So the output will heavily dependent on the corpus being used.

The snapshot in figure 2 shows the extracted Concepts above a certain threshold and their corresponding fuzzy context vector. For example, the Context vector of the concept 'IMAGE' being extracted is shown as

IMAGE, mri(0.536), resonance(0.409), magnetic(0.427), modality(0), biomedical(0.847), lundervold(0.507), piscataway(0.608), application(0.298), clinical(0.439), diagnosis(0.916), ultrasound(0.776), automatic(0.817), chen(0.615)

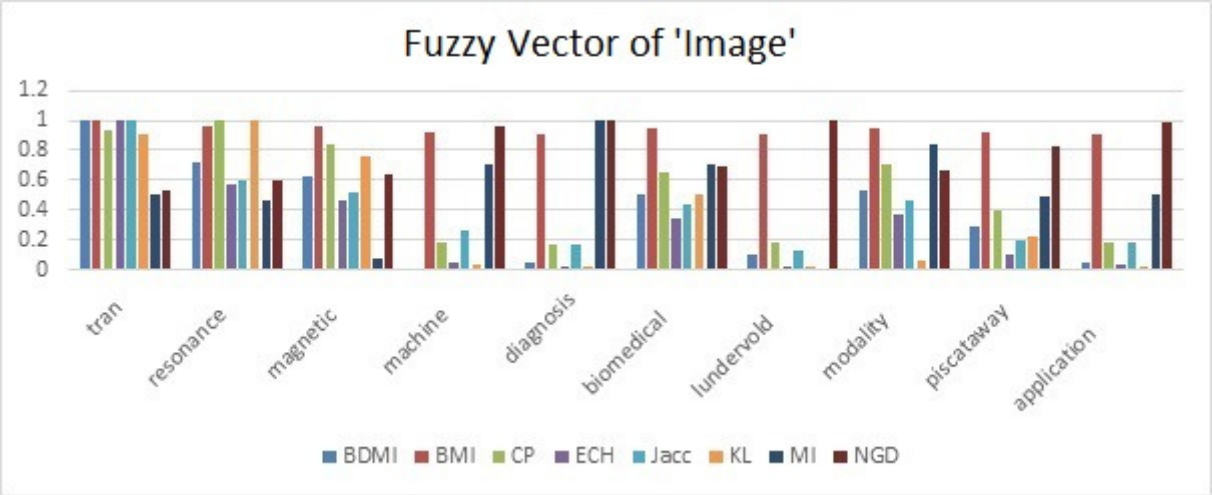

Figure 4. Fuzzy context vector extraction of term "Image" with different techniques

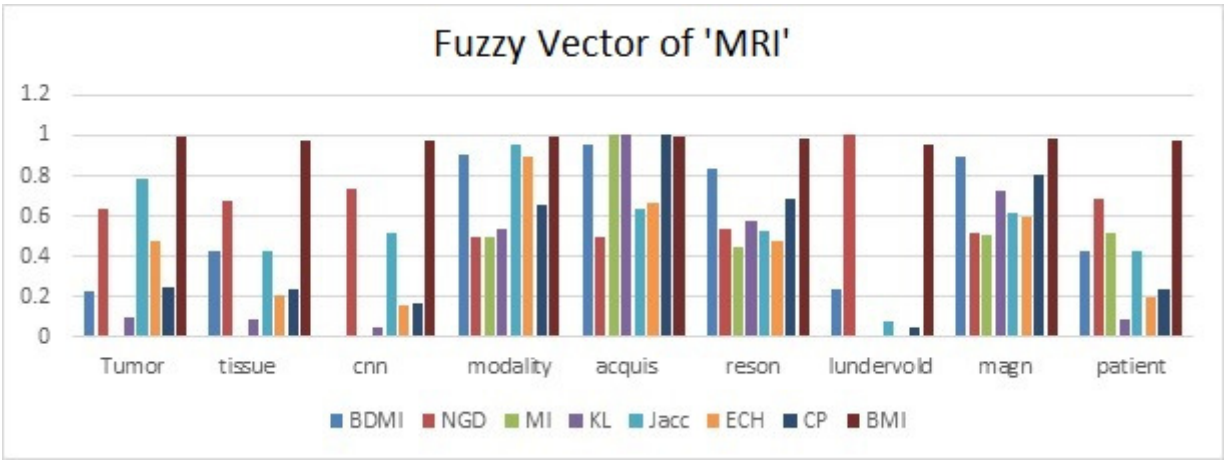

Figure 5. Fuzzy context vector extraction of term "MRI" with different techniques

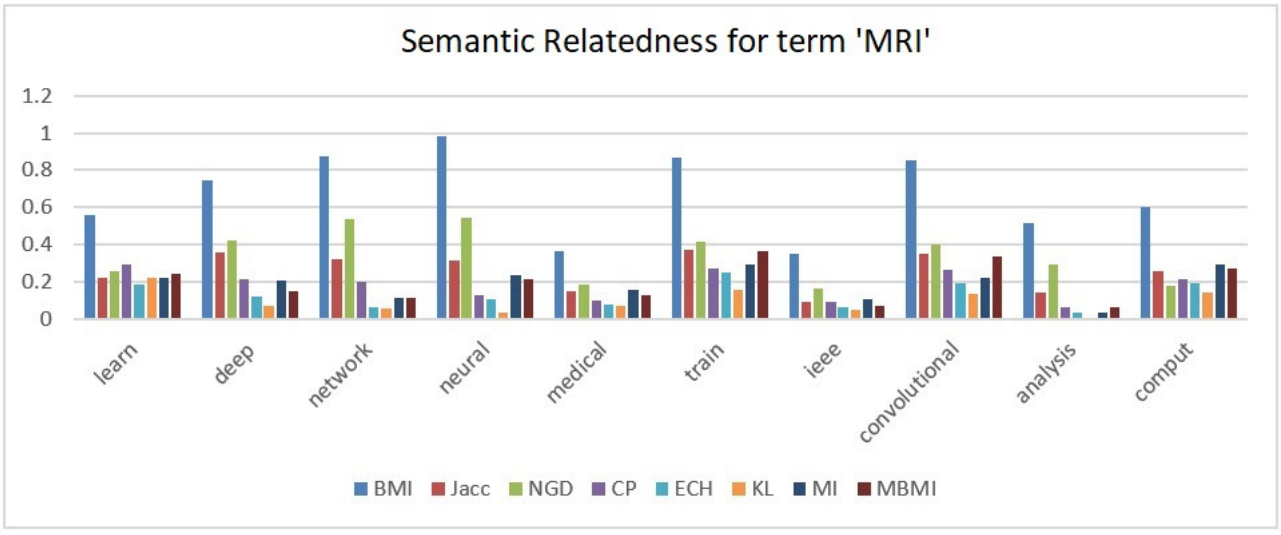

Figure 6. Semantic Similarity of term "MRI" with terms using different techniques

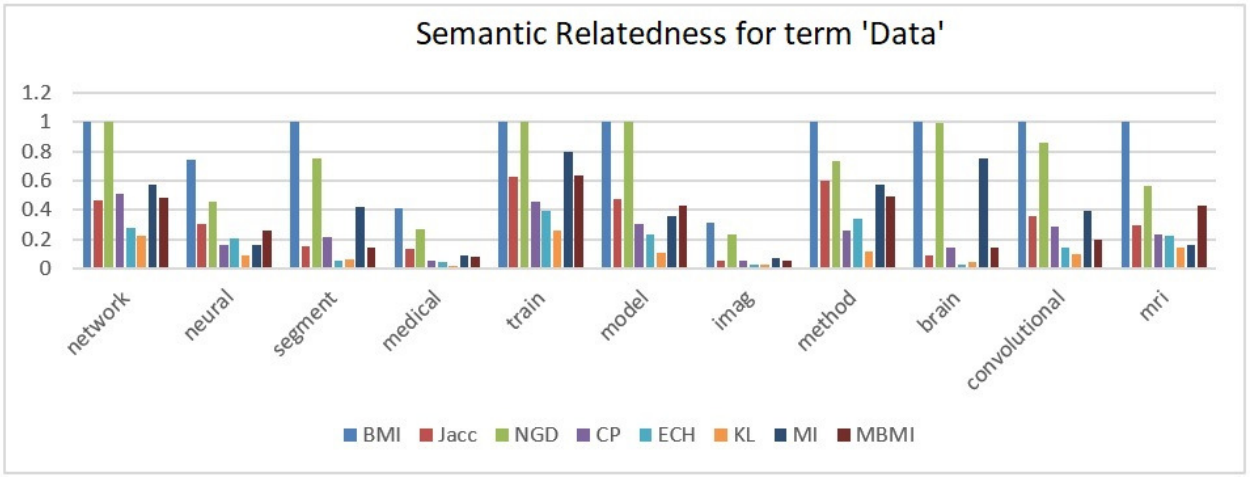

Figure 7. Semantic Similarity of term "DATA" with terms using different techniques

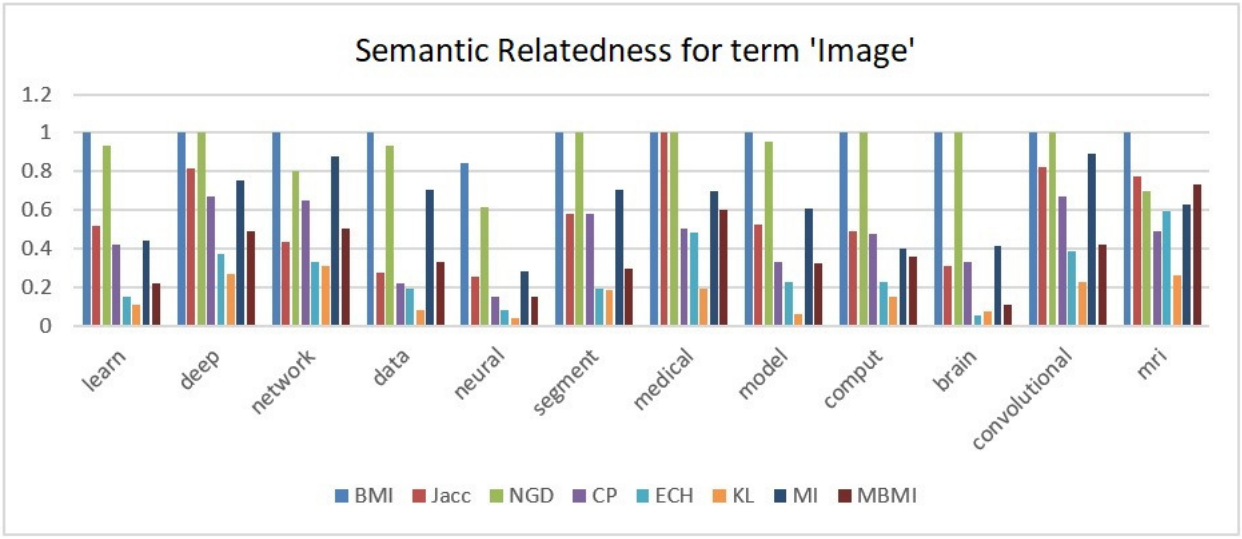

Figure 8. Semantic Similarity of term "Image" with terms using different techniques

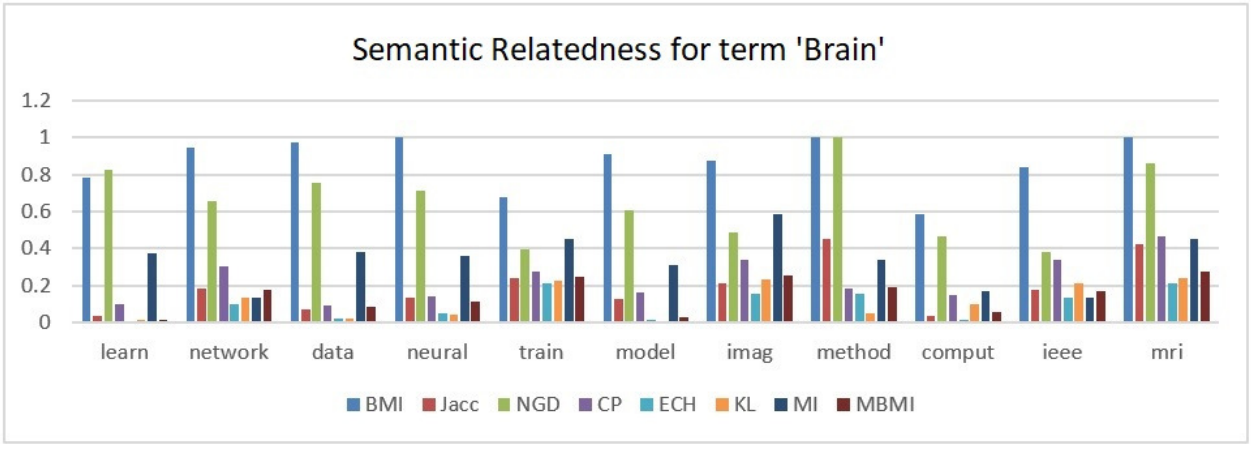

Figure 9. Semantic Similarity of term "Brain" with terms using different techniques

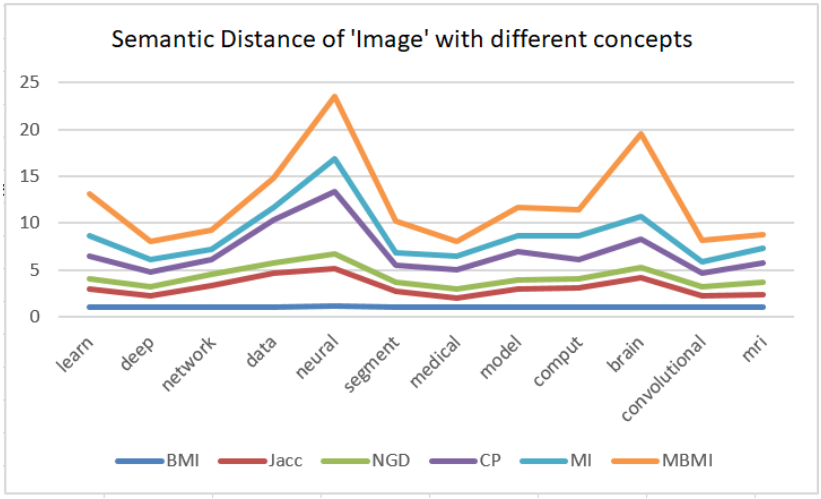

Figure 10. Semantic Distance of term "Image" with cross domain terms using different techniques

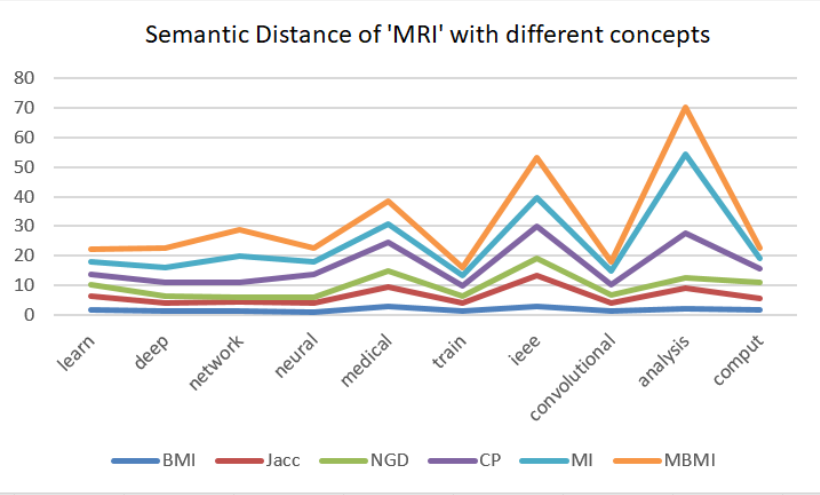

Figure 11. Semantic Distance of term "MRI" with cross domain terms using different techniques

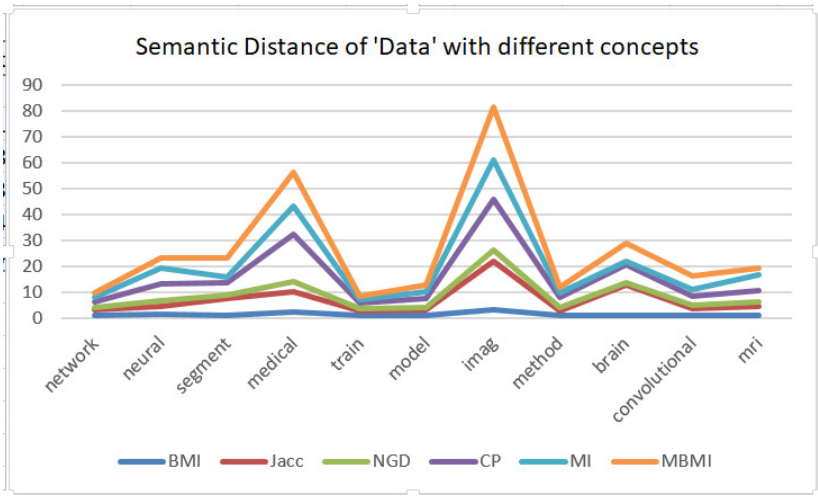

Figure 12. Semantic Distance of term "Data" with cross domain terms using different techniques

Table 1. Data Source Details

S. No	Reference	Word Count	Unique Word	Selected Concepts
1	Deep Learning and Its Applications in Biomedicine [38]	7024	2114	11
2	Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning[39]	4161	773	10
3	NiftyNet: a deep-learning platform for medical imaging[40]	12477	3149	13
4	An overview of deep learning in medical imaging focusing on MRI [41]	17291	2847	13
5	Deep Learning in Medical Ultrasound Analysis: A Review[42]	25985	3534	18
6	Review of MRI-based brain tumor image segmentation using deep learning methods[43]	15339	3334	8

The figure 4 and 5 shows the fuzzy context vector extracted for the concepts 'Image' and 'MRI'. The extracted Fuzzy Vector of These concepts has been discussed in the data set we have taken for our experiment as we can see that BMI gives a higher value for all elements in the vector, although it is not true whereas MBMI gives optimized value neither too low nor too high. The Figure 6,7, 8, and 9 and shows the semantic relatedness of concepts 'MRI' 'DATA', 'IMAGE' and 'BRAIN' with other concepts using different techniques like BMI, Jack, NGD, CP, ECH, KL, MI and our method MBMI. In figure 6, MRI is not related to deep and network, but BMI gives values 0.747 and 0.878, respectively, whereas MBMI gives values 0.152 and 0.115, respectively. In figure 7, DATA is related highly nearly 1 with other concepts using BMI technique, whereas it gives semantic similarity 0.144, 0.2, and 0.431 with concepts brain convolutional and MRI, respectively. In figure 8, Image is related with value 1 with almost all concepts, whereas MBMI gives similarity values 0.224, 0.492, 0.505, 0.329, and 0.151 with learn, deep, network and data respectively. In Figure 9, Brain is related with learn, network, data, method and IEEE with values 0.781, 0.947, 0.972 , 1 and 0.842 respectively using BMI techniques. Whereas using MBMI we get values 0.015, 0.178, 0.081, 0.189 and 0.17. In Figure 10 and 11, it can be shown that the very dissimilar terms are dissimilar shown by our existing technique as compared to other techniques. Thus we have seen that MBMI has been useful in the scenario when we have vast and multi-domain data i.e., Big Data having an unstructured part of being processed for e-learning. The result is dependent on the distributional probabilities of the terms, content, and domain also.

6. Conclusion and Future Work

A novel semantic relatedness technique has been proposed and experimented for semantic relatedness computation for cross-domain unstructured data. We have used academic articles related to medical journals. These articles are related to medical science and computer science. Our techniques give better results primarily when two concepts are not related altogether. However, the proposed work can also be done for finding similar and dissimilar authors based on their paper publication. It can be applied for the clustering of news articles, and fake news can be detected. Furthermore, a crawler can be made to download articles from Elsevier or other academic data sources and since processing textual data takes much time so Big Data tools like Hadoop or Spark can be used to process these articles, and relevant results can be obtained.

References

[1] Xindong Wu, Xingquan Zhu, Gong-Qing Wu, and Wei Ding. Data mining with big data. *IEEE transactions on knowledge and data engineering*, 26(1):97–107, 2013.

[2] Kiran Adnan and Rehan Akbar. An analytical study of information extraction from unstructured and multidimensional big data. *Journal of Big Data*, 6(1):91, 2019.

[3] Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407, 1990.

[4] George W Furnas, Scott Deerwester, Susan T Dumais, Thomas K Landauer, Richard A Harshman, Lynn A Streeter, and Karen E Lochbaum. Information retrieval using a singular value decomposition model of latent semantic structure. In *Proceedings of the 11th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 465–480. ACM, 1988.

[5] Mustapha Baziz, Mohand Boughanem, Nathalie Aussenac-Gilles, and Claude Chrisment. Semantic cores for representing documents in ir. In *Proceedings of the 2005 ACM symposium on Applied computing*, pages 1011–1017. ACM, 2005.

[6] Upali S Kohomban and Wee Sun Lee. Learning semantic classes for word sense disambiguation. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 34–41. Association for Computational Linguistics, 2005.

[7] Siddharth Patwardhan, Satanjeev Banerjee, and Ted Pedersen. Using measures of semantic relatedness for word sense disambiguation. In *International conference on intelligent text processing and computational linguistics*, pages 241–257. Springer, 2003.

[8] Simone Paolo Ponzetto and Michael Strube. Knowledge derived from wikipedia for computing semantic relatedness. *Journal of Artificial Intelligence Research*, 30:181–212, 2007.

[9] David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. Automatic evaluation of topic coherence. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 100–108. Association for Computational Linguistics, 2010.

[10] Mark Stevenson and Mark A Greenwood. A semantic approach to ie pattern induction. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 379–386. Association for Computational Linguistics, 2005.

[11] IA Budan and H Graeme. Evaluating wordnet-based measures of semantic distance. *Comutational Linguistics*, 32(1):13–47, 2006.

[12] Daqian Shi, Ting Wang, Hao Xing, and Hao Xu. A learning path recommendation model based on a multidimensional knowledge graph framework for e-learning. *Knowledge-Based Systems*, page 105618, 2020.

[13] Amir Hossein Nabizadeh, José Paulo Leal, Hamed N Rafsanjani, and Rajiv Ratn Shah. Learning path personalization and recommendation methods: A survey of the state-of-the-art. *Expert Systems with Applications*, page 113596, 2020.

[14] Ching Y Suen. N-gram statistics for natural language understanding and text processing. *IEEE transactions on pattern analysis and machine intelligence*, (2):164–172, 1979.

[15] Alberto Barrón-Cedeno, Paolo Rosso, Eneko Agirre, and Gorka Labaka. Plagiarism detection across distant language pairs. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 37–45. Association for Computational Linguistics, 2010.

[16] James L Peterson. Computer programs for detecting and correcting spelling errors. *Communications of the ACM*, 23(12):676–687, 1980.

[17] Patrick AV Hall and Geoff R Dowling. Approximate string matching. *ACM computing surveys (CSUR)*, 12(4): 381–402, 1980.

[18] Temple F Smith, Michael S Waterman, et al. Identification of common molecular subsequences. *Journal of molecular biology*, 147(1):195–197, 1981.

[19] Saul B Needleman and Christian D Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*, 48(3):443–453, 1970.

[20] Matthew A Jaro. Probabilistic linkage of large public health data files. *Statistics in medicine*, 14(5-7):491–498, 1995.

[21] William E Winkler. String comparator metrics and enhanced decision rules in the fellegi-sunter model of record linkage. 1990.

[22] Weixin Zeng, Xiang Zhao, Jiuyang Tang, Zhen Tan, and Xuqian Huang. Cleek: A chinese long-text corpus for entity linking. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 2026–2035, 2020.

[23] Raymond YK Lau, Dawei Song, Yuefeng Li, Terence CH Cheung, and Jin-Xing Hao. Toward a fuzzy domain ontology extraction method for adaptive e-learning. *IEEE transactions on knowledge and data engineering*, 21 (6):800–813, 2008.

[24] Paul Jaccard. Étude comparative de la distribution florale dans une portion des alpes et des jura. *Bull Soc Vaudoise Sci Nat*, 37:547–579, 1901.

[25] Rudi L Cilibrasi and Paul MB Vitanyi. The google similarity distance. *IEEE Transactions on knowledge and data engineering*, 19(3):370–383, 2007.

[26] Rada Mihalcea, Courtney Corley, Carlo Strapparava, et al. Corpus-based and knowledge-based measures of text semantic similarity. In *Aaai*, volume 6, pages 775–780, 2006.

[27] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[28] Jay J Jiang and David W Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. *arXiv preprint cmp-lg/9709008*, 1997.

[29] Philip Resnik. Using information content to evaluate semantic similarity in a taxonomy. *arXiv preprint cmp-lg/9511007*, 1995.

[30] Patrick Pantel and Dekang Lin. A statistical corpus-based term extractor. In *Conference of the Canadian Society for Computational Studies of Intelligence*, pages 36–46. Springer, 2001.

[31] Eric Yeh, Daniel Ramage, Christopher D Manning, Eneko Agirre, and Aitor Soroa. Wikiwalk: random walks on wikipedia for semantic relatedness. In *Proceedings of the 2009 workshop on graph-based methods for natural language processing*, pages 41–49. Association for Computational Linguistics, 2009.

[32] David Milne. Computing semantic relatedness using wikipedia link structure. In *Proceedings of the new zealand computer science research student conference*, 2007.

[33] Michael Strube and Simone Paolo Ponzetto. Wikirelate! computing semantic relatedness using wikipedia. In *AAAI*, volume 6, pages 1419–1424, 2006.

[34] Rafeeq Ahmed and Nesar Ahmad. Knowledge representation by concept mining & fuzzy relation from unstructured data. *published in International Journal of Research Review in engineering Science and Technology (ISSN 2278-6643) Volume-1 Issue-2*, 2012.

[35] Raymond YK Lau, Jin Xing Hao, Maolin Tang, and Xujuan Zhou. Towards context-sensitive domain ontology extraction. In *2007 40th Annual Hawaii International Conference on System Sciences (HICSS'07)*, pages 60–60. IEEE, 2007.

[36] Tanvir Ahmad, Rafeeq Ahmad, Sarah Masud, and Farheen Nilofer. Framework to extract context vectors from unstructured data using big data analytics. In *2016 Ninth International Conference on Contemporary Computing (IC3)*, pages 1–6. IEEE, 2016.

[37] Fernando Pech-May, Alicia Martinez-Rebollar, Jorge Magana-Govea, Luis A Lopez-Gomez, and Edna M Mil-Chontal. Semantic annotation approach for information search. *Research in Computing Science*, 148:59–73, 2019.

[38] Chensi Cao, Feng Liu, Hai Tan, Deshou Song, Wenjie Shu, Weizhong Li, Yiming Zhou, Xiaochen Bo, and Zhi Xie. Deep learning and its applications in biomedicine. *Genomics, proteomics & bioinformatics*, 16(1):17–32, 2018.

[39] Daniel S Kermany, Michael Goldbaum, Wenjia Cai, Carolina CS Valentim, Huiying Liang, Sally L Baxter, Alex McKeown, Ge Yang, Xiaokang Wu, Fangbing Yan, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*, 172(5): 1122–1131, 2018.

[40] Eli Gibson, Wenqi Li, Carole Sudre, Lucas Fidon, Dzhoshkun I Shakir, Guotai Wang, Zach Eaton-Rosen, Robert Gray, Tom Doel, Yipeng Hu, et al. Niftynet: a deep-learning platform for medical imaging. *Computer methods and programs in biomedicine*, 158:113–122, 2018.

[41] Alexander Selvikvåg Lundervold and Arvid Lundervold. An overview of deep learning in medical imaging focusing on mri. *Zeitschrift für Medizinische Physik*, 29 (2):102–127, 2019.

[42] Shengfeng Liu, Yi Wang, Xin Yang, Baiying Lei, Li Liu, Shawn Xiang Li, Dong Ni, and Tianfu Wang. Deep learning in medical ultrasound analysis: A review. *Engineering*, 2019.

[43] Ali Işın, Cem Direkoğlu, and Melike Şah. Review of mri-based brain tumor image segmentation using deep learning methods. *Procedia Computer Science*, 102:317–324, 2016.