Fuzzy Cross Domain Concept Mining

Rafeeq Ahmed, Tanvir Ahmad

***Abstract*: E-Learning has emerged as an important research area. Concept maps creation for emerging new domains such as e-Learning is even more challenging due to its ongoing development nature. For creating Concept map, concepts are extracted. Concepts are domain dependent but big data can have data from different domains. Data in different domain has different semantics. So before applying any analytics to such big unstructured data, we have to categorize the important concepts domain wise semantically before applying any machine learning algorithm. In this paper, we have used a novel approach to automatically cluster the E-Learning concept semantically; we have shown the cluster in table format. Initially, we have extracted important concepts from unstructured data followed by generation of vector space of each concept. Then we used different similarity formula to calculate fuzzy membership values of elements of vector to its corresponding concepts. Semantic Similarity is calculated between two concepts by considering repeatedly the semantic similarity or information gain between two elements of each vector. Then Semantic similarity between two concepts is calculated. Thus concept map can be generated for a particular domain. We have taken research articles as our dataset from different domains like computer science and medical domain containing articles on Cancer. A graph is generated to show that fuzzy relationship between them for all domain. Then clustering them in based on their distances**

Index Terms*: *Semantic Mining, Concept Map Extraction, Multidomain Mining, Text Mining.

I. INTRODUCTION

Data Analytics is done on all type of data including unstructured data. Unstructured data is defined as data without any scheme or any particular format or order. Also percentage of data in unstructured format is largest among all format of data i.e. structured, semi unstructured data, which is almost 80 percent of data in industry. Because of unstructured in nature, many different algorithms has been proposed which falls under text mining area. Text mining is done over emails, e-learning articles, web pages, news articles, commercial sites, product reviews, blogs and other forms of textual communication. Thus it can generate big revenue and has got industrial application like analysis of movie reviews, product reviews, emails, blogs, sentiment analysis, question answer generation, text summarization and so on. Text mining has been used for E-learning Concept Extraction, Information Retrieval/Extraction, Semantic Web, Named Entity Recognition (NER), Ontological based fields and many more.Text mining algorithms performs well for small data but as we know data is increasing day by day with an exponential rate and traditional software's fails to process Big Data in real time. That's why Big Data tools like Hadoop, Spark have been developed which works in distributed mode.

Data Analytics now a day refers to Big Data Analytics which has brought a big revolution in Computer Science as well as other fields. Big Data has been defined by some important properties like Volume, Variety, velocity, Value, Veracity. Variety includes means structured, semi structured, unstructured data stored together, processed, and analyzed. One more important property is that it has data with different domains like Wikipedia has pages representing different domains. Thus to process Big Data first we need to separate data according to different domains especially if we are looking for E-Learning topics where we want to generate concept or learning path, etc. So existing algorithms do not take care of this important feature of data.

So in this paper, we have worked for cross domain unstructured data which automatically separates data semantically we have taken research articles from medical and computer science as our data set. So we have used the technique of semantic relatedness computation to compute the similarity of two terms. Broadly there are four categories to compute similarity measures like String based similarity measures, Character based similarity measures, knowledge and Corpus based similarity measures which provides a strong foundation for research fields like text classification, information retrieval, topic detection, document clustering, topic tracking, text summarization , question answering system, essay scoring, neural machine translation, questions generation and others.

Text Similarity computation or semantic relatedness measures similarities between two words. Words of a natural language have similarity either semantically or lexically. This is the basic and starting task for computation of sentence, paragraph or document similarity, and followed by this machine learning algorithm can be applied for clustering which will also involve Singular value decomposition (SVD) [1], Principal Component Analysis (PCA), LSA (Latent Semantic Analysis) [2] and so on. Similarities in character sequence of two words referred as lexical similarity and semantically similar words have same thing or having same context or one is sub/super class of another. For lexical similarity, a String-Based algorithm works on the sequence/pattern of string and characters. It is a metric for measuring similarity/dissimilarity between words or string partially or completely. For getting Semantic similarity we have mainly Corpus-Based or Knowledge-Based algorithms. Former algorithm computes semantic similarity based on information gained from large data silos while later one extracts information from semantic networks. Application of Semantic relatedness is word sense disambiguation[3][4], spelling correction[5], or coreference resolution[6], performing semantic indexing for information retrieval[7], or assessing topic coherence[8], information extraction patterns[9]. The contribution of our

Revised Manuscript Received on July 22, 2019.

Rafeeq Ahmed, Computer Engg Dept, Jamia Millia Islamia, New Delhi, India.

Tanvir Ahmad, , Computer Engg Dept, Jamia Millia Islamia, New Delhi, India.

work is as follows:

- We have pointed out an important issue in property of Big Data that huge volume of data can belong to different domain in many cases.
- We have proposed a model for carrying out concept extraction of cross domain Big Data.
- Concepts which are learning topics belonging to different domains are automatically clustered according to their domain by using techniques of fuzzy membership formula.

Organization of our work is as follows: Section 2 discusses about semantic relatedness. Section 3 focuses on implementation and in section 4 results have been shown and discussed. Next section has conclusion and future work.

II. TEXT MINING

Similarity measures for text data are based either on String or knowledge or Corpus based similarity measures. String Based algorithms are further categorized into Characters based and Term based algorithms.

Characters based algorithms are Longest Common SubString (LCS), Damerau-Levenshtein [10][11], Jaro [12][13],Jaro–Winkler[14],Needleman-Wunschalgorithm[15], Smith-Waterman[16], N-gram[17].

Term-based Similarity Measures are boxcar distance, absolute value distance, L1 distance, Block Distance or commonly known as Manhattan distance [18], Cosine similarity, Dice's coefficient [19], Euclidean distance, Jaccard similarity [20], Matching Coefficient, Overlap coefficient.

Corpus-Based similarity includes Hyperspace Analogue to Language (HAL) [21][22], Latent Semantic Analysis (LSA)[2].

In general text mining involves extractions of terms, concept using Bag of Words (BOW) or N-Gram so that each document can be represented by vector space. Thus a matrix is created and machine learning algorithms are applied over here. If matrix columns are large then we apply technique of mathematics called Singular Value Decomposition is reduce number of features maintaining similarity structure among rows. Other techniques are Generalized Latent Semantic Analysis (GLSA) [23], Explicit Semantic Analysis (ESA)[24], The cross-language explicit semantic analysis (CLESA)[25], Pointwise Mutual Information - Information Retrieval(PMI-IR)[26], Second-order co-occurrence Pointwise mutual information (SCO-PMI) [27][28],Normalized Google Distance (NGD)[29].

$$NGD\ (t_i, t_j) = \frac{max\{log_2|w_{t_i}|,\ log_2|w_{t_i}|\} - log_2\left|w_{t_i,t_j}\right|}{log_2|w+1| - min\{log_2|w_{t_i}|,\ log_2|w_{t_i}|\}} \quad (1)$$

$$KL(t_i, t_j) = \sum Pr(t_i|t_j)\ log_2 \frac{Pr(t_i|t_j)}{Pr(t_i)} \quad (2)$$

$$ECH(t_i, t_j) = Pr\ (t_j) \sum Pr(t_i|t_j)\ log_2 \frac{Pr(t_i|t_j)}{Pr(t_i)} \quad (3)$$

Knowledge-Based Similarity is also an important semantic similarity measures for getting similarity between concepts or word or terms based on the information extracted from semantic networks [30]. WordNet [31] which has got large lexical database of English is most used semantic network. Besides these mentioned above, algorithms based on information content are Jiang and Conrath [32], Resnik [33] and Lin [34].

There are many algorithms developed to get semantic matching of words or sentences or topics or pages, even categories e.g. for Wikipedia pages, Wikirelate[35] based on category tree, Wikipedia Link Vector Model[36] using Wikipedia link structure, WikiWalks[37]. There are many algorithms developed to get corpus relevance or mutual information which is in general based on windowing process over the whole data set. The algorithms are conditional probability (CP), Kullback-Leibler divergence (KL), Expected Cross Entropy (ECH), Jaccard (JA), and Balanced Mutual Information (BMI) [38].

$$\begin{aligned} BMI\ (t_i, tj) \quad = \\ \beta \times [\ Pr(t_i, t_j)\ log_2\left(\frac{Pr(t_i, t_j)+1}{\boldsymbol{Pr(t_i)Pr(t_j)}}\right)\ + \\ Pr(\neg t_i, \neg t_j)\ log_2\left(\frac{Pr(\neg t_i, \neg t_j)+1}{Pr(\neg t_i)\ Pr(\neg t_j)}\right)\ ] \\ -\ (1-\beta) \times [\ Pr(t_i, \neg t_j)\ log_2\left(\frac{Pr(t_i, \neg t_j)+1}{Pr(t_i)\ Pr(\neg t_j)}\right) + \\ Pr(\neg t_i, t_j)\ log_2\left(\frac{Pr(\neg t_i, t_j)+1}{Pr(\neg t_i)\ Pr(t_j)}\right)\ ] \end{aligned} \quad (4)$$

In adaptive e-learning, Concept map is used as a support system for teachers to assist learners and by monitoring learner's performance on the fly by providing Remedial-Instruction Path (RIP) [39]. Concept map can be analyzed also as it was done in [40] generated by undergraduate students. Concept maps can be created from textual and non-textual sources, e.g. concept maps for Croatian language has been given in [41]. It's easy to understand real life problems or solution if it is presented in a graphical way and thus concept map or concept hierarchy are better tools to represent knowledge and organize them too. Thus author has focused on extraction concepts from textbooks using the knowledge in Wikipedia pages [42].

Concept to concept distance based on WordNet senses and WordNet Topic mapping to WordNet Domain has been done in [43]. Extending to this, the both type of associations can also be calculated among concepts. This has been done and evaluated in [44] where they used two datasets direct and indirect associations between two concepts has also been calculated and evaluated on two datasets 'gene-disease' association dataset and 'disease-symptom-treatment' associations. The datasets they used are DisGeNET public database and "MedicineNet.net" website respectively.

Automatic construction of concept map by using text-mining techniques for e-Learning domain has been done in [45][46].

So they used a small set of research articles including

conferences and journals as their main data source in e-Learning domain. Concept map is useful for knowledge representation. Fuzzy Context vector has been used to represent a concept [47]. Again this can has been extended with big data analytics in [48].

III. IMPLEMENTATION

In E-learning, as we know that an academic articles for any particular concepts describes that concepts frequently thus from statistics from point of view, frequency of that concept will be higher. Thus we had made certain assumptions that each frequent word or term above threshold will certainly represent an essential concept. Concept is basic entity for construction of concept or any knowledge representation methodology, domain ontology, semantic networks, Ontology, model, theory and so on [49]

Now each concept has a lexical meaning as well as semantic meaning. We are mainly focus on semantics of that concept. Semantics can be extracted by co-occurring words or terms in sentence or documents. There will be some relation between concept and co-occurring words that'swhy, they been present there.

Collocation expressions [50] can be explained in the sense of a group of similar words semantically, original words can be highly related to adjacent words in an expression unit [51].

A. Data preprocessing

Tokenization: Tokens are extracted after preprocessing of unstructured document. Tokens are basic entity which represents a phrase, word, symbol or other events using delimiters to identify them in the source document.

Stop Word Elimination: In unstructured data there is no statistical importance of stop words but they are important in discovering opinion, sentiments analysis, and events detection, mainly in natural language processing. In natural Language processing noun noun, noun adjective etc. can extracted using tools like parts of speech tagger.

Stemming: Words with inflexion and common morphologic lings are stemmed up to their base or root form. This process is known as stemming. Stemming can be done by using Porter stemmer [52]

B. Concept Extraction

After data preprocessing steps, tokenization, removing stop words, noisy data, stemming, and the tokens in a documents are stored in a list, then frequency is calculated. Frequent words above certain threshold are stored as concept for which fuzzy vector space will be generated, which is explained in the following sub sections.

C. Concept's Fuzzy Vector Extraction

Distributional hypothesis explains that two concept or topic reveals similarity directly proportional to sharing of similar linguistics contexts. Collocation terms are group of semantically similar words because of the probability of their co-occurrence is high. Statistical information among tokens or concepts can be extracted using their distributional probability in the documents. Thus we have extracted context vector for every concept in a document by doing windowing process [53]. Windowing process starts with the scanning of documents by taking a virtual window of δ words and if the concept is present in that window then other present words in that window are also added to the context vector of that concept if not already present in the context vector.

$$Concept\ C_i = \{t_1, t_2 \dots t_m\} \quad (5)$$

To generate fuzzy vector space or fuzzy context vector, there exists several algorithms which are used to obtain the mutual relationship between two terms.

Semantic relatedness between two entities can also be computed to by following formula [2]:

$$MI(t_i, t_j) = log_2\left(\frac{Pr(t_i, t_j)}{Pr(t_i) * Pr(t_j)}\right) \quad (6)$$

$$R(t_i, t_j) = log_2\left(\frac{w_{ij}(t_i, t_j) * w}{w(t_i) * w(t_j)}\right) \quad (7)$$

$$BMI\ (t_i, t_j) = \beta \times \left[\frac{w_{ij}}{w} log_2\left(\frac{(w_{ij}+w)\times w}{w_i \times w_j}\right) + \frac{(w-w_{ij})}{w} log_2\left(\frac{(2w-w_{ij})\times w}{(w-w_i)\times(w-w_j)}\right)\right] -(1-\beta)\times \left[\frac{(w_i-w_{ij})}{w} log_2\left(\frac{(w_i-w_{ij}+w)\times w}{w_i \times (w-w_j)}\right) + \frac{(w_j-w_{ij})}{w} log_2\left(\frac{(w_j-w_{ij}+w)\times w}{(w-w_i)\times w_j}\right)\right] \quad (8)$$

In fig 1, we have taken N no of documents, then we have done tokenization, taking each token we have done stemming then checked for stop word. It is added to the list if not already present in the list, otherwise incremented the count. In fig 2, we are calculating mutual information between two terms by doing the windowing process. We have already stored the frequency of each term, we have to calculate the probability of occurrence of two terms together in window, and i.e. we have to obtain the total no of window containing both term t_1 and t_2. In fig 3, the flowchart is showing the steps to extract the fuzzy context vector for all important concepts. Every element in the vector has some fuzzy membership value.

Thus a context vector with fuzzy set i.e. fuzzy vector space for a concept C_i with terms t_i, having membership function μ_{c_i} can be defined as

$$Concept\ C_i = \{t_1\ (\mu_{c_i}(t_i)),\ t_2\ (\mu_2) \dots t_m\ (\mu_m)\} \quad (9)$$

Here ti is semantically related term with the concept which is extracted through windowing process and corresponding weight is μ_{c_i} computation of which is described below:

$$\mu_i = \frac{Fr\ (T_i/C_j)}{Fr\ (C_j)} \quad (10)$$

D. Semantic Relatedness Computation

The semantic relatedness among concepts can be obtained by using intensive computation. High value means two concepts are highly related and vice versa. Following this, distance among concepts has been computed.

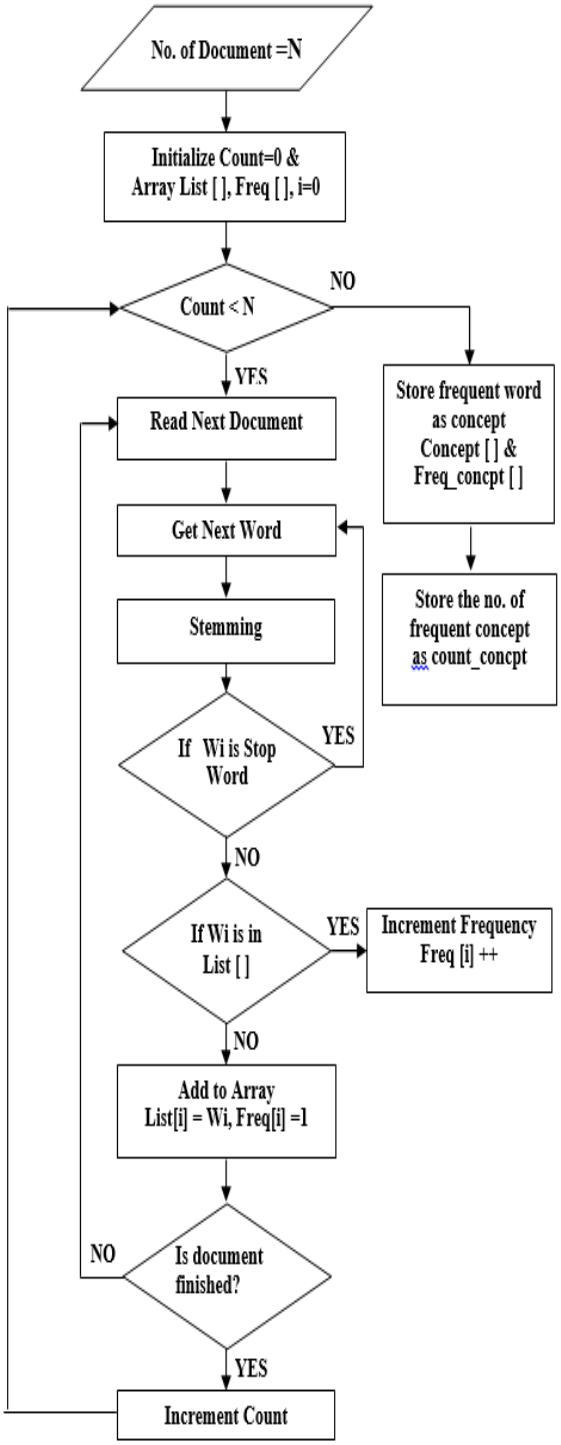

Fig 1: Concept Extraction

Let the two context vector be suppose context vector be C_1 and C_2 with fuzzy context vector with n terms as

$$C_1 = \{t_1\,(\mu_{c_1}(t_1)),\ t_2\,(\mu_{c_1}(t_2)\ \ldots\ t_n\,(\mu_{c_1}(t_n))\ \}\ \&$$

$$C_2 = \{t_1\,(\mu_{c_2}(t_1)),\ \ t_2\,(\mu_{c_2}(t_2)\ \ldots\ t_n\,(\mu_{c_2}(t_n))\ \}$$

$$R\left(t_{(i}, c_{1)}\,, t_{(m(i),c_2)}\right) = \max_{0\le j\le n}\left\{MI\left(t_{i_{c_1}}, t_{j_{c_2}}\right)\right\} \quad (11)$$

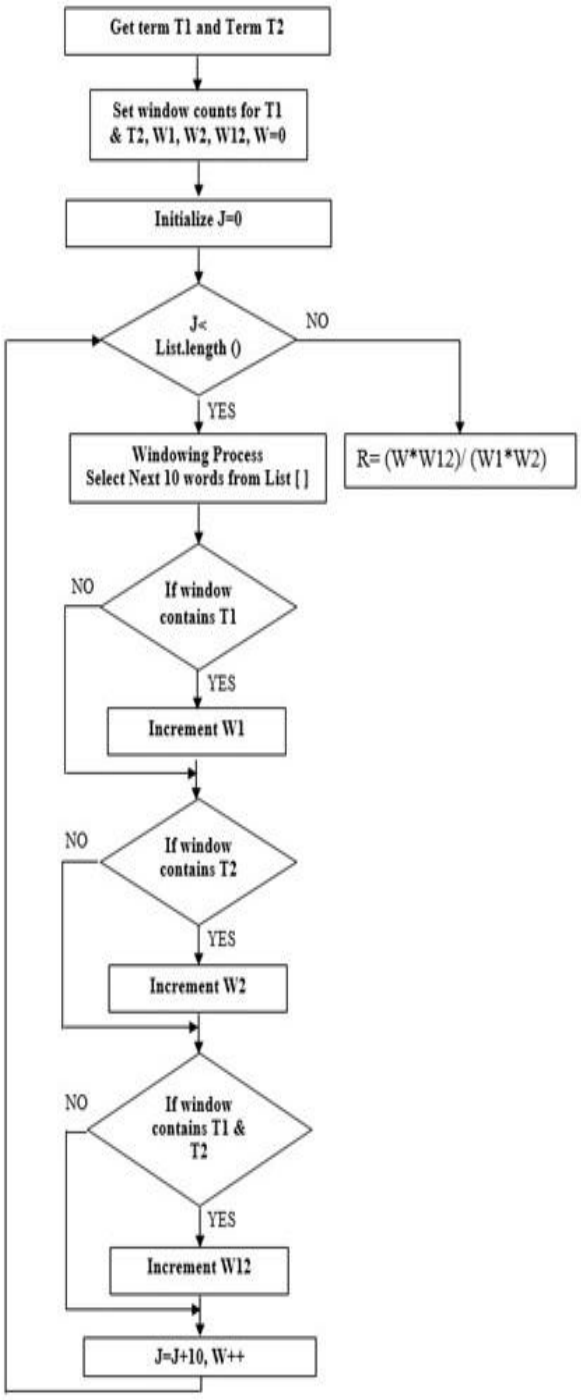

Fig2: Mutual Information Computation by windowing Process

where m(i) stores the index of term j of second vector for which term i has highest semantic relatedness for i=1to n. Now context distance is computed as follows

$$Dist(c_1, c_2) = \frac{1}{n}\sum_{i=1}^{n}\Big(R\big(t_i, t_{m(i)}\big) \times \mu_{c_i}(t_i) \times \mu_{c_2}\big(t_{m(i)}\big)\ \Big) \quad (12)$$

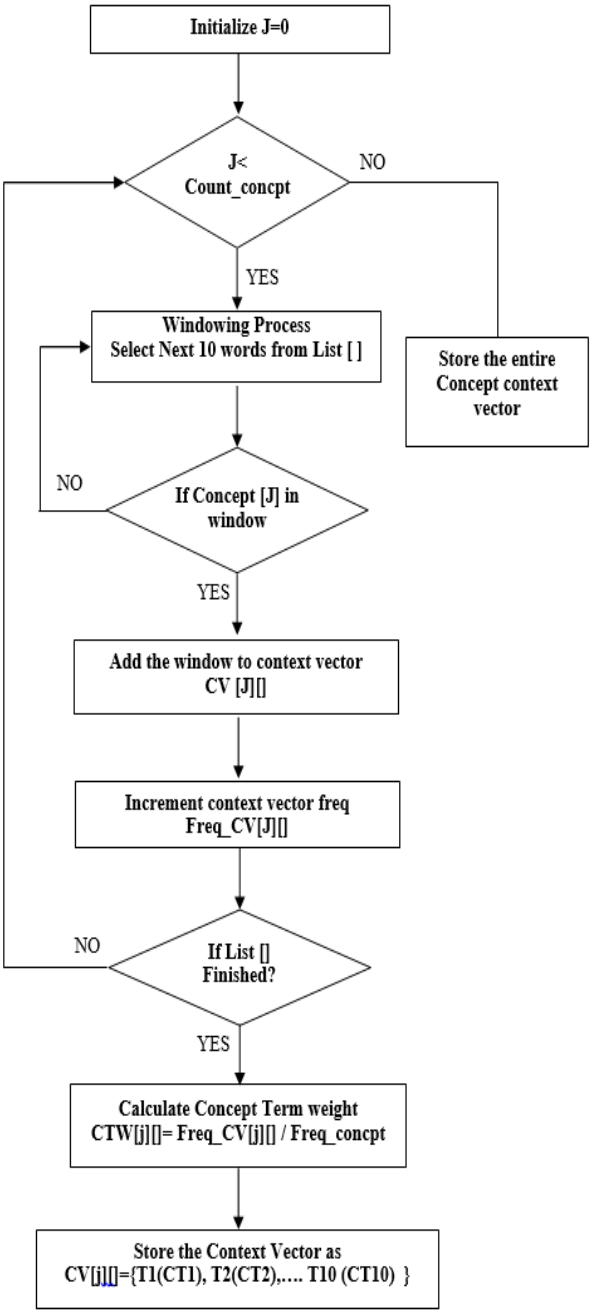

Fig 3:Fuzzy Context Vector Generation

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

In Figure 4, we have shown the steps to get semantic relatedness between two concept using their fuzzy vector. First we selected two concepts C_i and C_j, then for term t_i in C_i, we have done windowing process for each term t_j present in C_j, then we selected term with maximum match i.e. selected the $t_{m(i)}$, we will store the product of these two, and weights too. Similarly, we will do for all terms of C_i with that of C_j and add up the products to get distance between Concept C_i and C_j divided by the length of the vector.

IV. RESULTS

We have taken data set is the content of research paper on ontology and medical articles related to cancer stored all together. These are unstructured data. As we know that Big Data contains huge amount of data and obviously for unstructured it will different data from different domains. Like E-Learning articles if we consider it, they are having different domains all together and if we want to apply analytics on these literatures there is no work which can process unstructured data from different domains all together.

In Table 1 and 2 we have shown a table for different values of Beta. The elements of the vector for a concept will be fixed but fuzzy membership will vary depending on different functions. Depending on the beta we can say the vector is changing as per values of β. What beta says that it controls the impact on semantic relatedness between two concepts if both concepts are absent in a window.

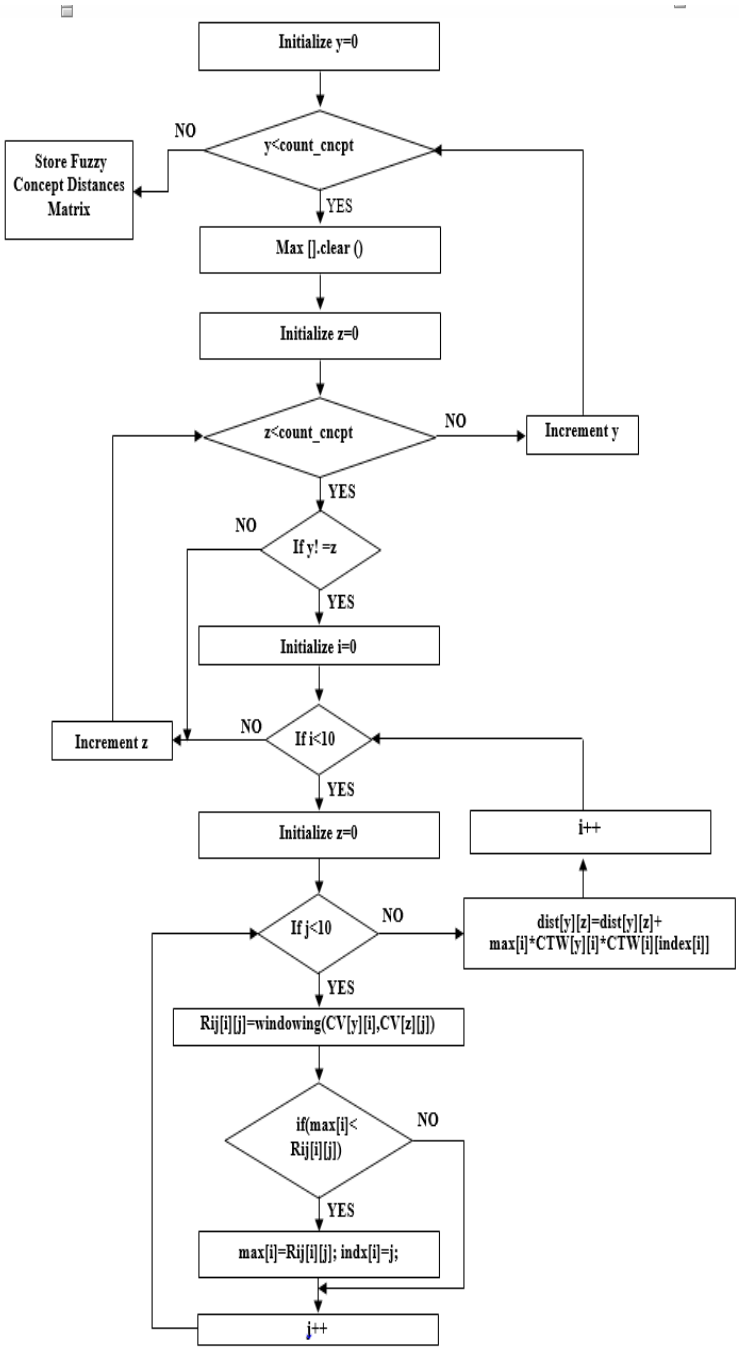

Fig 4: Context Distance between two Fuzzy Concept

Thus fig 5 refers to the membership values of elements of vector for the concept Ontology. Now Fig 6 &7 refers to the semantic relatedness between two concepts consideringβ as parameter. We can see for the value of β=0.35 gives the relevance if two terms are r present together or absent together.

Table 3 shows the clustering of different concepts based on fuzzy membership value which successfully does clustering of related concepts together.

A. Complexity Analysis

As the Dimensionality Reduction works in a high dimensional environment with large datasets, complexity of the reduction or clustering techniques is a major issue. Hence for a term-document matrix, the complexity of Dimensionality Reduction varies based on the reduction technique used. For the Integrated Feature Selection Method, the Time Complexity is $O(N^2)$.Individual feature evaluation Focusing on identifying relevant features without handling feature redundancy Time complexity: $O(N^2)$Identifying Relevant Features Relying on minimum feature subset heuristics to implicitly handling redundancy while pursuing relevant features.

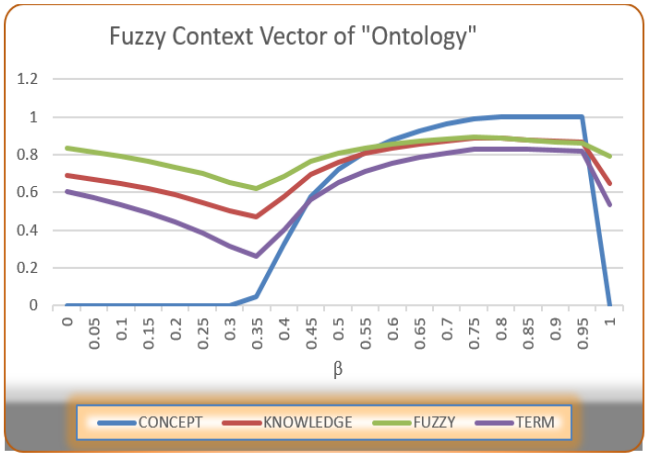

Fig 5: Fuzzy membership with concept "Ontology"

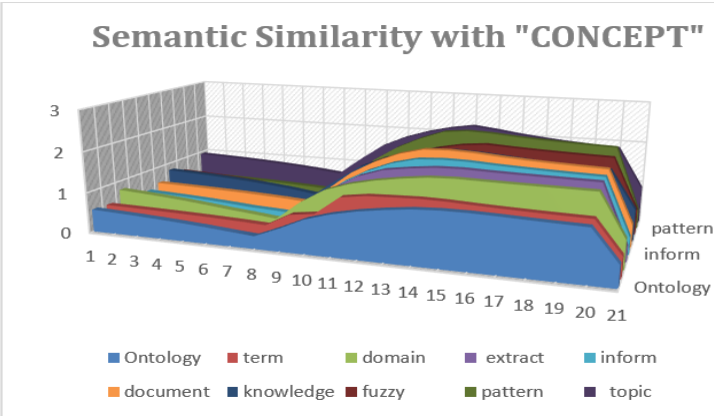

Fig 6 Similarity of different concept with "Concept"

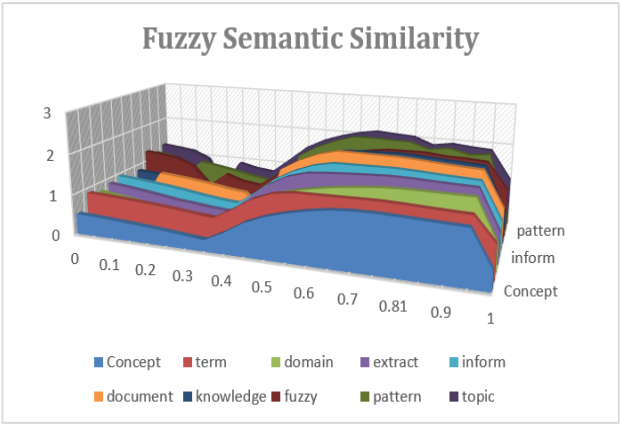

Fig 7: Similarity of different concept with "Concept"

Table 1: Ontology with different Beta

β	Ontology	term	domain	extract	inform	document	knowledge	fuzzy	pattern	topic
0	0.552	0.527	0.779	0.437	0.476	0.58	0.816	0.675	0.488	0.942
0.05	0.529	0.519	0.744	0.406	0.453	0.567	0.791	0.648	0.478	0.922
0.1	0.504	0.509	0.703	0.369	0.426	0.553	0.762	0.615	0.467	0.898
0.15	0.474	0.497	0.656	0.324	0.394	0.536	0.727	0.575	0.453	0.871
0.2	0.44	0.483	0.599	0.278	0.356	0.517	0.686	0.519	0.437	0.837
0.25	0.4	0.465	0.537	0.261	0.31	0.492	0.635	0.428	0.416	0.796
0.3	0.352	0.444	0.484	0.24	0.285	0.463	0.572	0.35	0.391	0.745
0.35	0.307	0.415	0.422	0.215	0.255	0.425	0.492	0.337	0.357	0.686
0.4	0.548	0.724	0.876	0.698	0.706	0.819	0.918	0.784	0.787	1.129
0.45	0.816	0.8	1.224	1.157	1.164	1.275	1.307	1.189	1.277	1.534
0.5	0.99	1.241	1.419	1.48	1.485	1.601	1.568	1.464	1.642	1.799
0.55	1.112	1.324	1.549	1.674	1.718	1.839	1.755	1.661	1.916	1.987
0.6	1.201	1.345	1.641	1.756	1.859	1.985	1.879	1.809	2.128	2.128
0.65	1.27	1.362	1.711	1.817	1.892	2.002	1.891	1.923	2.181	2.221
0.7	1.299	1.35	1.732	1.83	1.882	1.976	1.864	1.976	2.158	2.151
0.75	1.301	1.321	1.72	1.81	1.845	1.926	1.815	1.94	2.107	2.065
0.8	1.285	1.298	1.711	1.795	1.817	1.887	1.766	1.911	2.068	1.999
0.85	1.263	1.28	1.705	1.784	1.794	1.857	1.725	1.888	2.038	1.946
0.9	1.245	1.266	1.699	1.774	1.776	1.833	1.693	1.869	2.013	1.904
0.95	1.23	1.254	1.695	1.766	1.761	1.813	1.666	1.853	1.993	1.869
1	0.504	0.509	0.703	0.369	0.426	0.553	0.762	0.615	0.467	0.898

V. CONCLUSION & FUTURE WORK

This work is basically based on unstructured data where we have taken data set as journals from computer science and medical area. Cross domains mining is done to extract concept which we treat them as the E-learning topic. Furthermore we find their semantic vector through windowing process and then extracted fuzzy vector based on semantics by using similarity formula. Then we find relations among concept using these fuzzy values and we have shown that concept in different domains are automatically clustered into their domain. Formula for similarity measures work well for small data i.e. they will fail if data becomes too large. Even using Big Data tools like Hadoop or Spark, these formulas then also will not give

relevant value because of Big Data characteristics. So these formula needs to be updated for Big Unstructured Multi Domain Data.

Table2 for "Concept" with different Beta

β	Concept	term	domain	extract	inform	document	knowledge	fuzzy	pattern	topic
0	0.5	0.885	0.79	0.872	0.941	0.56	0.852	1.28	1.144	1.267
0.05	0.479	0.864	0.735	0.823	0.902	0.24	0.816	1.227	1.1	1.218
0.1	0.455	0.839	0.672	0.767	0.857	0.988	0.775	1.168	0.5	1.162
0.15	0.428	0.811	0.601	0.701	0.805	0.945	0.728	0.98	0.993	0.97
0.2	0.396	0.777	0.52	0.633	0.744	0.895	0.673	0.5	0.925	0.22
0.25	0.359	0.737	0.437	0.59	0.673	0.836	0.609	0.853	0.846	0.934
0.3	0.316	0.688	0.377	0.539	0.615	0.763	0.534	0.706	0.75	0.828
0.35	0.278	0.649	0.343	0.506	0.579	0.711	0.481	0.647	0.68	0.775
0.4	0.518	0.908	0.749	0.896	0.911	0.19	0.877	0.956	0.25	1.143
0.45	0.804	1.198	0.73	1.268	1.281	1.413	1.275	1.278	1.429	1.522
0.5	0.997	1.384	1.259	1.516	1.532	1.685	1.553	1.488	1.701	1.76
0.55	1.134	1.449	1.383	1.65	1.713	1.881	1.756	1.635	1.895	1.925
0.6	1.236	1.455	1.472	1.691	1.812	1.994	1.893	1.743	2.041	2.047
0.65	1.314	1.461	1.539	1.721	1.817	1.985	1.915	1.826	2.047	2.125
0.7	1.35	1.468	1.591	1.744	1.822	1.978	1.932	1.89	2.03	2.088
0.75	1.356	1.474	1.632	1.763	1.826	1.973	1.945	1.897	2.018	2.059
0.81	1.344	1.459	1.644	1.754	1.805	1.943	1.918	1.902	1.871	1.876
0.85	1.323	1.436	1.64	1.733	1.774	1.904	1.879	1.845	1.938	1.952
0.9	1.307	1.417	1.637	1.716	1.749	1.873	1.848	1.837	1.819	1.902
0.95	1.293	1.402	1.635	1.702	1.728	1.848	1.822	1.798	1.873	1.87
1	0.455	0.839	0.672	0.767	0.857	0.988	0.775	1.168	0.5	1.162

Table3: Clustering of learning concept semantically

Cluster 0	Cluster 1	Cluster 2
CONCEPT	CELL	TERM
ONTOLOGY	CANCER	WORD
LEARN	TUMOR	BAS
DOMAIN	GENE	DOCUMENT
EXTRACT	HUMAN	SET
INFORM	STUDY	PATTERN
KNOWLEDGE	GROWTH	TOPIC
FUZZY	EXPRESS	SYSTEM
RELATION	BREAST	SEMANTIC
TECHNIQUE	GENOME	FORM
METHOD	IDENTIFY	RESULT
MODEL	FUNCT	FOUND
REPRESENT	MIRNA	WEB
EXAMPLE	FACTOR	SELECT
DATA	ASSOCIAT	APPROACH
RESEARCH	PROTEIN	MIN
DEFIN	ENDOTHELI	PROCES
SECT	ACTIN	TEXT
HIERARCHY	TARGET	ADD
STATISTIC		DETECT
ALGORITHM		

REFERENCES

1. Furnas, George W., et al. "Information retrieval using a singular value decomposition model of latent semantic structure." *Proceedings of the 11th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 1988.
2. Landauer, Thomas K., and Susan T. Dumais. "The latent semantic analysis theory of acquisition." *Induction and Representation of Knowledge1997* (1997).
3. Patwardhan, Siddharth, Satanjeev Banerjee, and Ted Pedersen. "Using measures of semantic relatedness for word sense disambiguation." *International conference on intelligent text processing and computational linguistics*. Springer, Berlin, Heidelberg, 2003.
4. Kohomban, Upali S., and Wee Sun Lee. "Learning semantic classes for word sense disambiguation." *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 2005.
5. Budan, I. A., and H. Graeme. "Evaluating wordnet-based measures of semantic distance." *Comutational Linguistics*32.1 (2006): 13-47.
6. Ponzetto, Simone Paolo, and Michael Strube. "Knowledge derived from Wikipedia for computing semantic relatedness." *Journal of Artificial Intelligence Research* 30 (2007): 181-212.
7. Baziz, Mustapha, et al. "Semantic cores for representing documents in IR." *Proceedings of the 2005 ACM symposium on Applied computing*. ACM, 2005.
8. Newman, David, et al. "Automatic evaluation of topic coherence." *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 2010.
9. Stevenson, Mark, and Mark A. Greenwood. "A semantic approach to IE pattern induction." *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 2005.
10. Hall, Patrick AV, and Geoff R. Dowling. "Approximate string matching." *ACM computing surveys (CSUR)* 12.4 (1980): 381-402.
11. Peterson, James L. "Computer programs for detecting and correcting spelling errors." *Communications of the ACM* 23.12 (1980): 676-687.
12. Advances in record linkage methodology as applied to the 1985 census of Tampa Florida
13. Jaro, Matthew A. "Probabilistic linkage of large public health data files." *Statistics in medicine* 14.5-7 (1995): 491-498.
14. Winkler, William E. "String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage." (1990).
15. Needleman, Saul B., 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 (1970): 443-453.
16. Smith, Temple F., and Michael S. Waterman. "Identification of common molecular subsequences." *Journal of molecular biology* 147.1 (1981): 195-197.
17. Barrón-Cedeno, Alberto, et al. "Plagiarism detection across distant language pairs." *Proceedings of the 23rd International Conference on Computational Linguistics*. Association for Computational Linguistics, 2010.
18. F.K Eugene, "Taxicab Geometry" ,Dover. ISBN 0-486-25202-7.
19. Dice, Lee R. "Measures of the amount of ecologic association between species." *Ecology* 26.3 (1945): 297-302.
20. P. Jaccard, Etude comparative de la distribution florale dans une portion des Alpes et des Jura, Bulletin de Societe Vaudoise des Sciences Naturelles 37, 547-579.
21. Lund, Kevin. "Semantic and associative priming in high-dimensional semantic space." *Proc. of the 17th Annual conferences of the Cognitive Science Society, 1995*. 1995.
22. Lund, Kevin, and Curt Burgess. "Producing high-dimensional semantic spaces from lexical co-occurrence." *Behavior research methods, instruments, & computers* 28.2 (1996): 203-208.
23. Matveeva, Irina, et al. "Generalized latent semantic analysis for term representation." *Proc. of RANLP*. 2005.
24. Gabrilovich, Evgeniy, and Shaul Markovitch. "Computing semantic relatedness using wikipedia-based explicit semantic analysis." *IJcAI*. Vol. 7. 2007.
25. Potthast, Martin, Benno Stein, and Maik Anderka. "A Wikipedia-based multilingual retrieval model." *European conference on information retrieval*. Springer, Berlin, Heidelberg, 2008.
26. Turney, Peter D. "Mining the web for synonyms: PMI-IR versus LSA on TOEFL." *European conference on machine learning*. Springer, Berlin, Heidelberg, 2001.

27. Islam, Aminul, and Diana Inkpen. "Semantic text similarity using corpus-based word similarity and string similarity." *ACM Transactions on Knowledge Discovery from Data (TKDD)* 2.2 (2008): 10.
28. Islam, Aminul, and Diana Inkpen. "Second Order Co-occurrence PMI for Determining the Semantic Similarity of Words." *LREC*. 2006.
29. Cilibrasi, Rudi L., and Paul MB Vitanyi. "The google similarity distance." *IEEE Transactions on knowledge and data engineering* 19.3 (2007): 370-383.
30. Mihalcea, Rada, Courtney Corley, and Carlo Strapparava. "Corpus-based and knowledge-based measures of text semantic similarity." *AAAI*. Vol. 6. 2006.
31. Miller, George A. "WordNet: a lexical database for English." *Communications of the ACM* 38.11 (1995): 39-41.
32. Jiang, Jay J., and David W. Conrath. "Semantic similarity based on corpus statistics and lexical taxonomy." *arXiv preprint cmp-lg/9709008* (1997).
33. Resnik, Philip. "Using information content to evaluate semantic similarity in a taxonomy." *arXiv preprint cmp-lg/9511007* (1995).
34. Pantel, Patrick, and Dekang Lin. "A statistical corpus-based term extractor." *Conference of the Canadian Society for Computational Studies of Intelligence*. Springer, Berlin, Heidelberg, 2001.
35. Strube, Michael, and Simone Paolo Ponzetto. "WikiRelate! Computing semantic relatedness using Wikipedia." *AAAI*. Vol. 6. 2006.
36. Milne, David. "Computing semantic relatedness using wikipedia link structure." *Proceedings of the new zealand computer science research student conference*. 2007.
37. Yeh, Eric, et al. "WikiWalk: random walks on Wikipedia for semantic relatedness." *Proceedings of the 2009 Workshop on Graph-based Methods for Natural Language Processing*. Association for Computational Linguistics, 2009.
38. Lau, Raymond YK, et al. "Toward a fuzzy domain ontology extraction method for adaptive e-learning." *IEEE transactions on knowledge and data engineering* 21.6 (2009): 800-813.
39. Lee, Chun-Hsiung, Gwo-Guang Lee, and Yungho Leu. "Application of automatically constructed concept map of learning to conceptual diagnosis of e-learning." *Expert Systems with Applications* 36.2 (2009): 1675-1684.
40. Sengul, Sare, and S. Can Senay. "Assessment of Concept Maps Generated by Undergraduate Students about the Function Concept." *Procedia-Social and Behavioral Sciences*116 (2014): 729-733.
41. Žubrinić, Krunoslav. "Automatic creation of a concept map." (2011).
42. Wang, Shuting, et al. "Concept hierarchy extraction from textbooks." *Proceedings of the 2015 ACM Symposium on Document Engineering*. ACM, 2015.
43. Gella, Spandana, Carlo Strapparava, and Vivi Nastase. "Mapping WordNet Domains, WordNet Topics and Wikipedia Categories to Generate Multilingual Domain Specific Resources." *LREC*. 2014.
44. Sadoddin, Reza, and Osvaldo Driollet. "Mining and Visualizing Associations of Concepts on a Large-Scale Unstructured Data." *2016 IEEE Second International Conference on Big Data Computing Service and Applications (BigDataService)*. IEEE, 2016.
45. Chen, Nian-Shing, Chun-Wang Wei, and Hong-Jhe Chen. "Mining e-Learning domain concept map from academic articles." *Computers & Education* 50.3 (2008): 1009-1021.
46. Malik, Hasmat, et al. "Applications of Artificial Intelligence Techniques in Engineering." *SIGMA* 1 (2018).
47. Ahmed, Rafeeq, 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).
48. Ahmad, Tanvir, et al. "Framework to extract context vectors from unstructured data using big data analytics." *2016 Ninth International Conference on Contemporary Computing (IC3)*. IEEE, 2016.
49. Zadeh, Lotfi A. "Fuzzy sets." *Information and control* 8.3 (1965): 338-353.
50. Perrin, Patrick, and Frederick E. Petry. "Extraction and representation of contextual information for knowledge discovery in texts." *Information sciences* 151 (2003): 125-152.
51. Chen, Nian-Shing, Chun-Wang Wei, and Hong-Jhe Chen. "Mining e-Learning domain concept map from academic articles." *Computers & Education* 50.3 (2008): 1009-1021.
52. Porter, Martin F. "An algorithm for suffix stripping." *Program*14.3 (1980): 130-137.
53. Lau, Raymond YK, et al. "Towards context-sensitive domain ontology extraction." *2007 40th Annual Hawaii International Conference on System Sciences (HICSS'07)*. IEEE, 2007.

AUTHORS PROFILE

Rafeeq Ahmed, research scholar of Jamia Millia Islamia, has done B.Tech (Computer Engineering) and M.Tech (Software Engineering) from Aligarh Muslim University. He has been given Gold medal in M.Tech. He also been awarded Maulana Azad National Fellowship (MANF). He has teaching experience of more than 4 years. He has published 9 International Journal and Conferences papers. Awarded. He has also got best paper award in international Conference SIGMA-2018 held at NSIT, New Delhi.

Dr. Tanvir Ahmad is currently Professor & Head of Computer Engineering Department as well as Additional Director of FTK-Centre for Information Technology, Jamia Millia Islamia. His area of research is Text Mining, Graph Mining, Big Data Analytics, Natural Language Processing, Information Security. He has teaching experience of more than 20 Years. He has supervised 6 PhD students and more than 20 students are doing research work under him. Number of papers indexed in SCI/SCIE/ SCOPUS : 40+, Number of Papers indexed in Google scholar : 54+. He has done MTech (IT) from IP University, Delhi and B.E.(CSE) from Bangalore University. He is Member of Advisory Group for Digital Locker system, National e-Governance Division, Ministry of I.T. and Communication, Govt. of India. He is member, Senate, Indian Institute of Information Technology (IIIT) Sonepat. He is a member of Expert Committee of National Board of Accreditation (NBA), India.

.