A Hybrid Approach for Measuring Semantic Similarity between Documents and its Application in Mining the Knowledge Repositories

Ms. K. L. Sumathy,
Research and Development center, Bharathiyar University,
Coimbatore,
Tamil Nadu

Dr. Chidambaram,
Asst prof, Rajah serofiji College,
Thanjavur,
Tamil Nadu

***Abstract*—This paper explains about similarity measure and the relationship between the knowledge repositories. This paper also describes the significance of document similarity measures, algorithms and to which type of text it can be applied Document similarity measures are of full text similarity, paragraph similarity, sentence similarity, semantic similarity, structural similarity and statistical measures. Two different frameworks had been proposed in this paper, one for measuring document to document similarity and the other model which measures similarity between documents to multiple documents. These two proposed models can use any one of the similarity measures in implementation aspect, which is been put forth for further research.**

Keywords—dataset documents; research similarity documents; ontology and corpus

I. INTRODUCTION

Objectives:

Now-a-days information on the web is increasing rapidly day-by-day. The increase of web based information and number of internet users', difficult to find the relevant documents for users to particular needs. In order to resolve this issue this paper proposes semantic similarity based document retrieval.

Methods/Statistical Analysis:

The dataset documents are stored in the knowledge repositories. To get the relevant documents mining approach is used. The number of repeated words in a document considers as a keyword. Through preprocessing technique the repeated words in the documents will be removing. The term frequency mechanism proposes to identify important keywords in the documents. Term frequency determines the frequently occurring words in a document as keywords. The Jaccard similarity coefficient estimates the similarity between the documents and the ontology plays an important role in retrieval of similarity document.

Findings:

In retrieval process the pos tagger applied to the document and obtains noun, verb and adverb. This fed in to word net. Through word net the related keyword like synonyms, antonyms, and hypernyms are obtained. The SWETO technique provides similar documents from the knowledge repositories.

Accurately measuring semantic similarity between text documents presents a significant challenge due to the complexity and ambiguity of natural language semantics. Several natural language applications such as information retrieval, information recommendation, and machine translation require the similarity between sentences or documents. Generally, a pair of sentences or documents are said to be similar if they are predicted to have same meaning or conveys the same idea or subject. In natural language, there are different meanings in granularities such as word, phrase, sentence, and document. Word is the minimum mining unit, whereas sentence is the minimum unit to communicate some complete meaning. Moreover, there are various levels of similarities in natural languages. Words are generally categorized into synonyms and antonyms depending on the similarity between words and phrases. The calculation of similarities between documents is the basis for text classification and clustering. The techniques for similarity calculation vary in different levels. The word level similarity can be calculated from the spelling of words or the meaning of words. Word similarity is of two types such as symbolic similarity and semantic similarity. The symbolic similarity of words can be measured using the edit distance measure. The semantic similarity of words can be measured using WordNet.

The similarities between words in different sentences have a great impact on the similarity between two sentences. Words and their orders in the sentences are two chief factors to calculate sentence similarity. Sentence similarity is similar to the word similarity and document similarity. If words in two sentences are similar, the two sentences are said to be similar. Similarly, if sentences in two documents are similar, then the two documents are said to be similar. The sentence similarity measure considers the relation between words. The word similarity measures cannot calculate the sentence similarity as the word similarity reflects the closeness of two discrete words or concepts while the sentence similarity reflects the closeness of two sequences of words.

The similarities between sentences have a great impact on the similarity between documents. Most of the existing approaches calculate the document similarity based on the similarity between the keyword sets or the vectors of

keywords. Generally documents are represented in the form of bag-of-words while the meanings of documents are represented as vectors. The document similarity can be calculated using the cosine of the vectors. If the weight of the words is ignored, the document similarity can be measured based on the keywords set using Dice similarity or Jaccard Coefficient similarity.

A. Significance of Sentence/Document Semantic Similarity

Several recent applications of natural language processing demand an effective approach to calculating the similarity between sentences as in [1]. The deployment of sentence similarity can simplify the agent's knowledge base using natural sentences instead of using structural patterns of sentences. Semantic Similarity evaluates the similarity between concepts that are not lexicographically similar. The deep understanding of these concepts is necessary for computing semantic measures and for web mining. Similarity and relatedness measure can be applied to solve many problems in different applications. The measure of similarity and relatedness can be extended to many types of entities, such as words, sentences, texts, concepts, or Ontologies depending on the requirement. Lexical Semantics extracts semantic relations. Tasks such as document classification and clustering, information retrieval, and synonym extraction require precise measurement of semantic similarity between words. As the several applications and domains require semantic similarity, the measurement of sentence / document similarity has greater significance.

B. Potential Applications of Sentence/Document Semantic Similarity

Calculating semantic similarity among entities has application in several areas such as recommendation systems, e-commerce, search engines, biomedical informatics and in natural language processing tasks such as word sense disambiguation. In particular, user-based collaborative filtering tries to find people with similar tastes and recommend items to their peers liked by most of the people. The content-based recommender systems and search engines find items that are more similar to user queries. The sentence similarity has proven to be one of the effective techniques for enhancing retrieval performances as in [2]. The use of sentence representing the images can achieve a higher retrieval precision during the image retrieval from the web as in [3]. In text mining, sentence similarity act as an important factor to discover unseen knowledge from textual databases as in [4]. Semantic similarity efficiently evaluates the web search method of finding and ranking results. Hence, semantic similarity becomes vital in search engines as in [5]. Moreover, the short text similarity is important in applications like text summarization as in [6], text categorization as in [7], and machine translation as in [8].

II. Related Work

This section divides the related works into three parts such as sentence/document similarity based on metrics, sentence/document similarity based on methodologies, and hybrid approaches.

A. Sentence/Document Similarity Based on Metrics

There are several metrics for identifying the similarity between sentences as shown in table 1. The Jaccard similarity coefficient is measured by comparing the size of the intersection of words in two sentences with the size of the union of the words in two sentences as in [9]. The proportion of words that appear in both sentences normalized by the length of the sentence provides simple word overlap fraction as in [10]. The proportion of words that appear in two sentences weighted by their inverse document frequency is the value of IDF overlap [10]. The Zipfian overlap represents the Zipfian relationship between the length of words and their frequencies in a text collection as in [11]. IDF overlap is measured based on the sum of the product of term frequency and IDF of words that appear in both sentences as in [12]. The sums of IDF in the words that appear in both sentences are normalized by the overall lengths of the sentences and the relative frequency of words between two sentences as in [13].

TABLE I. Sentence/Document Similarity based on Metrics

Title	Similarity level	Metrics	Pros	Cons
Jacob B, et. al (2008) [9]	Sentence similarity	Jaccard similarity coefficient	Calculation of sentence similarity involves fewer computations	Word co-occurrence may be null. Considers only the surface similarity which is not reliable
Metzler. D et. al (2005) [10]		Simple word overlap fraction		
Metzler. D et. al (2005) [10]		IDF overlap		
Banerjee, S (2003) [11]		Zipfian overlap		
Allan. J et. al (2003) [12]		TF-IDF measure	Better performance	It requires a large text corpus for statistics computation
Hoad. T et.al (2003) [13]		Identity measure		
Chukfong H et.al (2003) [14]		WSD based measure		
Tversky, A (1977) [15]	Document similarity	Tversky's Contrast Model	These models are relatively simple	Natural language allows users to express the same meaning using different words and hence, calculating text similarity by word's surface similarity is implausible
Shepard. R et.al (1979) [16]		Common Features Model		
Rohde. D et.al (2002) [17]		Distinctive Features based contrast Model		
Lee. M et.al (2002) [18]		Local and global weighting functions	Weighted corpus representation improves the performance	It is hard to obtain large text corpus for a specific domain

Yun.M et.al (2013) [19] and Rushdi S et.al (2010) [20]		Conceptual tree similarity measure (WordNet Based)	Simple computation	This kind of similarity computation is based on literal similarity which is not highly accurate

B. Sentence/Document Similarity Based On Methodologies

The measurement of sentence/document similarity based on methodologies are classified into corpus-based statistical approaches, lexical based semantic similarity approaches, lexical based semantic similarity, ontologies based semantic similarity, relational based semantic similarity. Corpus-based Statistical Approaches are used to find the similarity between terms based on the corpus. Ontology is a significant resource for measuring the semantic similarity and relatedness. OWL represents the domain knowledge and plays a vital role in the application area of Artificial Intelligence (AI). The Relational based approaches compute the similarity based on the relation between the words.

C. Hybrid Approaches

The hybrid approach combines the semantic, corpus, ontology and relational based approaches. A novel hybrid approach extracts semantic knowledge from the structural representation of Wordnet and the statistic information on the Internet [32]. Internet based semantic knowledge estimates the semantic similarity between the two concepts in Wordnet. A useful measure called Normalized Google Distance (NGD) computes the semantic distance between the adjacent concepts, along the shortest path in WordNet using Internet semantic knowledge. It is one of the best approaches due to the deployment of Internet knowledge in WordNet based semantic relatedness measure.

TABLE II. HYBRID APPROACHES

Title	Features combination	Pros	Cons
Emiliano Giovannetti et.al (2008) [31]	Statistical methods and lexico-syntactic patterns	Improved accuracy	Data sparseness of the corpus
Liu et.al (2011) [32]	Structural, semantic network (Wordnet) and the Internet	Deployment of Internet knowledge improves overall system performance	
Jay J. Jiang and David W. Conrath (1997) [33]	Corpus statistics and lexical taxonomy	Useful in word sense disambiguation	

III. PROPOSED WORK

A. Combining Ontology Based And Count Based Similarity Model For Measuring Document Similarity

This work proposes a hybrid approach for measuring semantic similarity between documents. There is an extensive literature on measuring the similarity between the words, but there is less work related to the measurement of similarity between sentences and documents. This work measures the similarity between the documents using both ontology-based similarity model and counts based similarity model. Figure 1 shows the components used in this methodology. Initially, the proposed system represents the input documents as a bag of words, and it avoids the repeated terms. The pre-processing step removes stop words from the representation. The important keywords in the documents are identified using the term frequency mechanism. Term frequency determines the frequently occurring words in a document as keywords. The derived keywords are given to the ontology to obtain related keywords. The related keywords are given to the Jaccard similarity coefficient to decide the documents are similar or not. The Jaccard similarity coefficient is the count based similarity measure estimates the similarity between the documents by dividing the number of commonly related keywords by the total number of total keywords. In Jaccard coefficient result a value of "0" indicates the documents are completely dissimilar, "1" indicates that they are identical, and values between 0 and 1 represent a degree of similarity.

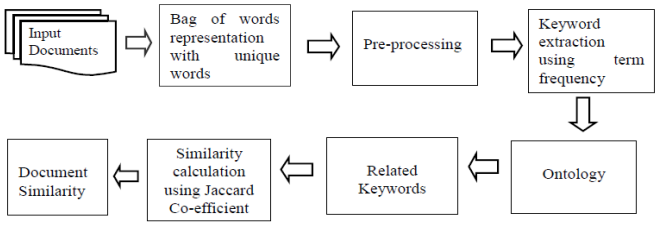

Fig. 1. Hybrid approach for measuring document similarity using ontology based and count based similarity model

B. Preprocessing

There is uncertainty of spelling convention in the language of Vietnamese. The majority typical ones contain "y" or "i" and situate marks on the syllables. Thus, preprocessing step intends to normalize information for additional investigation by one standard. Furthermore, this step intends to discover original words in named entities form like company's names, factoids, people, etc. We utilize usual expression as major technique to distinguish named entities.

C. Term Frequency Computation

Term frequency (TF) signifies number of concept occurrence in documents. Through text processing the concepts are extracted that characterized as vectors, the vectors calculated with Term Frequency. TF count up for documents facilitates to obtain elevated accuracy rate.

D. Ontology Construction

Ontology signifies the knowledge as set of concept. The Swoogle web search engine is also known as semantic engine. From the Swoogle OWL file (knowledge source) is extracted for ontology construction in Ontograph form by protégé to determine semantic weight is considered by discovering minimum distance for every concept in constructed ontograph. The automobile fragmentation ontograph is exposed below.

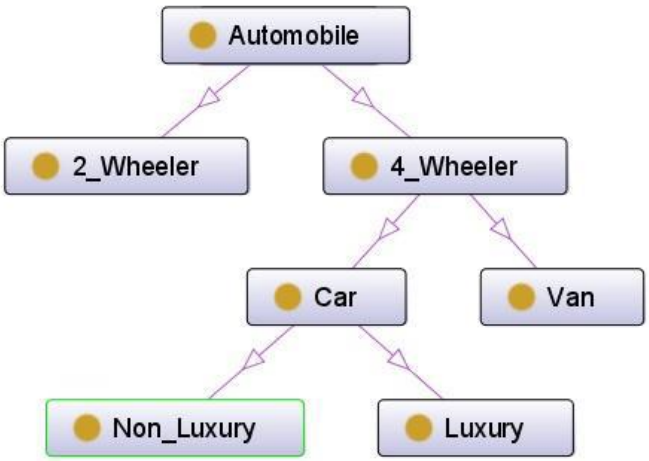

Fig. 2. Construction of Onto Graph

E. Jaccard Similarity Coefficient

Jaccard similarity is statistical measurement of the similarity among sample sets. Jaccard similarity coefficient is also known as Jaccard index. For both sets, it is defined as intersection of the cardinality divided by their cardinality union. Mathematically,

$$J(X, Y) = \frac{|X \cap Y|}{|X \cup Y|}$$

For Wikipedia data, slightly various approach required. For entities pair, computation is achieved on entities set with which pair elements occurred. For instance, in manipulating the pages similarity pairs, numerator is no. of .users that reduced both pages, no. of users who have reduced both or either in the denominator. Mathematically,

$$J(X, Y) = \frac{|A \cap B|}{|A \cup B|}$$

Where A and B match to the entities sets that take place with X and Y, correspondingly.

F. Hybrid Approach For Measuring Document Similarity Using Ontology And Corpus

The proposed system presents the hybrid framework to measure the similarity between the documents using WordNet ontology and Wikipedia corpus. Figure 3 explains the procedure involved in the proposed stem. Most of the existing approaches does not consider the document context in semantic similarity. It may lead to the inaccurate similarity measurement. Due to overcome this limitation the proposed approach exploits the Wikipedia and WordNet to identify the context of the document since, it has been extensively and effectively exploited to facilitate better understanding of documents. Moreover, both Wikipedia and WordNet are domain independent while they provide extensive coverage of almost every branch of knowledge. Initially, the input documents are applied to the POS tagger to obtain only nouns, verbs and adjectives from the input documents. The resultant terms of the POS tagger are fed into the WordNet ontology. It gives all the related keywords such as synonyms, antonyms, and hypernyms for a given input terms. The context of the each input documents is identified through the Wikipedia related articles for the related keywords. If the identified concepts from the Wikipedia for connecting document D1 with D2 are related then, the documents are considered as similar documents otherwise the documents are dissimilar.

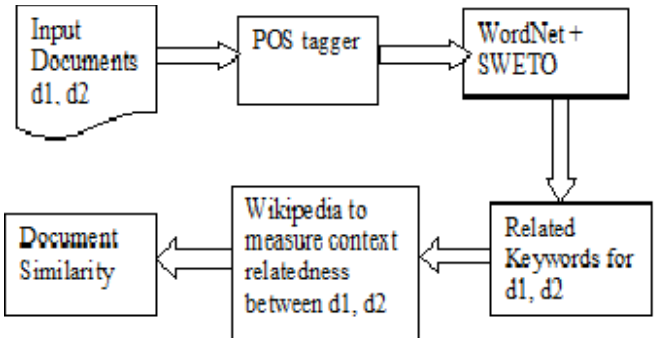

Fig. 3. Hybrid approach for measuring document similarity using ontology and corpus

G. POS Tagger Using Hidden Markov Model

HMM (Hidden Markov Model) is the statistical representation that is utilized to establish hidden parameter derived from observed parameters. It's extensively used, particularly in POS tagging for the input sequence.

Hidden part: T sequence tag
observed part: Word sequence
Transition probability:
aj-1,j = P(tsj|tsj-1) by hypothesis Markov-1, or *aj-1,j = P(tsj|tsj-1, tsj-2)* by hypothesis Markov-2
Output probability*: bj = P(wj|tsj)*

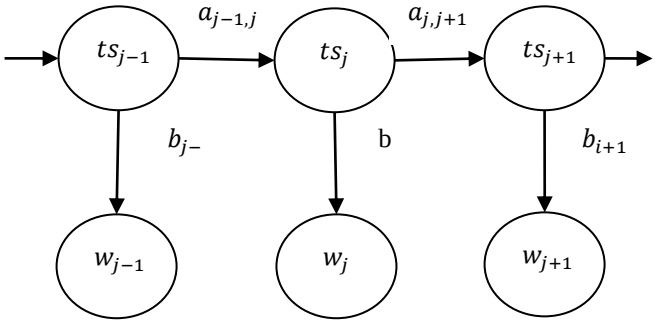

Fig. 4. POS tagging for HMM

The tagged sequence $\hat{T} = ts_1, ts_2, \ldots, ts_n$ suits

$$\hat{T} = \underset{T}{\operatorname{argmax}} P(w_1, w_2, \ldots, w_n | ts_1, ts_2, \ldots, ts_n) * P(ts_1, ts_2, \ldots, ts_n)$$

If we identify parts of speech of the word easily can determine word. The probability P(W|T) depends only basic probability $P(W_j|ts_j)$

$$P(W|T) \approx \prod_{j=1}^{n} P(W_j|ts_j)$$

Further computation for P(T)

$$P(T) = P(ts_1) * P(ts_2|ts_1) * \ldots\ldots\ldots * P(ts_n|ts_1, ts_2, \ldots, ts_n)$$

When we know previous POS probability we can predict POS probability appearing in sequence by Concern Markov-1 hypothesis. That means

$$P(T) = \prod_{j=1}^{n} P(ts_j | ts_{j-1})$$

Finally we obtain

$$\hat{T} = \underset{T}{\text{argmax}} \prod_{j=1}^{n} P(W_j | ts_j) * P(ts_j | ts_{j-1})$$

Where probabilities $P(W_j|ts_j)$and P $(ts_i|ts_{j-1})$ can be calculate by annotated corpus which is based on Maximum likelihood technique.

In similar manner, when concern Markov-2 hypothesis we obtain

$$\hat{T} = \underset{T}{\text{argmax}} \prod_{j=1}^{n} P(W_j | ts_j) * P(ts_j | ts_{j-2,} ts_{j-1})$$

Therefore, we can apply dynamic programming method Viterbi to resolve POS tagging.

H. Wordnet Ontology

Model descriptions of the word "country" on the WordNet seem like the subsequent. "people who be alive in the country or nation". Here Synset is a country and Hypernym is the nation. Hyponym is the people.

I. Sweto Ontology

SWETO Ontology is introduced by Large Scale Distributed Information Systems (LSDIS). Three SWETO versions are there namely small, medium and large.

Similarity among diverse ontologies concepts in following equations,

cj,ck are refer to the concepts

P()→ probability function.

$$\text{Mutual Similarity} = \frac{\log(p(c_j)p(c_k))}{P(c_j, c_k)}$$

$$P(c_j) = \frac{W_c}{W}$$

P (cj,ck) is a common terms joint probability distribution incident on same window and P (cj) is particular keyword ki probability appears in text window. The text window is text sequences frame in the web documents. To determine obscure concepts ambiguity we require membership of fuzzy which is related to mutual similarity. Let the function of fuzzy membership μi be in j^{th} concept and constant α and their value is being set at 0.5.

$$\mu_i(c_k) = \alpha \text{ x } P(c_i, c_k)\ log_2\left(\frac{P(c_j)P(c_k)}{P(c_j, c_k)}\right)$$

J. Information Content Based Measure

Both depth and path length relative measure utilize the information exclusively incarcerate with ontology to additionally establish the similarity among concept. In this sector the knowledge discovered with corpus is utilized to enhance information already currently in the taxonomy or ontologies. The approach of content based information is being referred the approach based on theory and corpus approach.

IV. Result and Discussion

A. Experimental Requirements

This section explains the materials required to evaluate the effectiveness of the proposed methodology.

B. Dataset

For document similarity, the dataset used is the Michael D.Lee document dataset, a collection of 50 documents from the Australian Broadcasting Corporation's news mail service. These documents were paired in all possible ways, and each of the 1,225 pairs has 8-12 human judgments.

C. Software Requirements

- Platform: Java
- IDE: Netbeans
- Database: MySQL Server
- Tool: Weka

D. Performance Metrics

- **Precision:** It is a ratio of correctly predicted similar documents to the total input documents.

$$\boldsymbol{Precision} = \frac{\text{Correctly predicted similar documents (TP)}}{\text{Total input documents (TP + FP)}}$$

- **Recall:** It is a ratio of correctly predicted similar documents to all similar sentences.

$$\boldsymbol{Recall} = \frac{\text{Correctly predicted similar documents (TP)}}{\text{Total input documents (TP + FN)}}$$

F-measure: It is a uniform harmonic mean of precision and recall.

$$\boldsymbol{Recall} = (1 + \beta)PR/(\beta P + R) = 2PR / (P + R)$$

(β=1, when precision and recall have the same weight)

Where,

TP: Number of documents predicted to be similar documents that actually are similar.

TN: Number of documents predicted to be dissimilar documents that actually are dissimilar

FP: Number of documents predicted to be similar that are actually dissimilar

FN: Number of documents predicted to be dissimilar that are actually similar

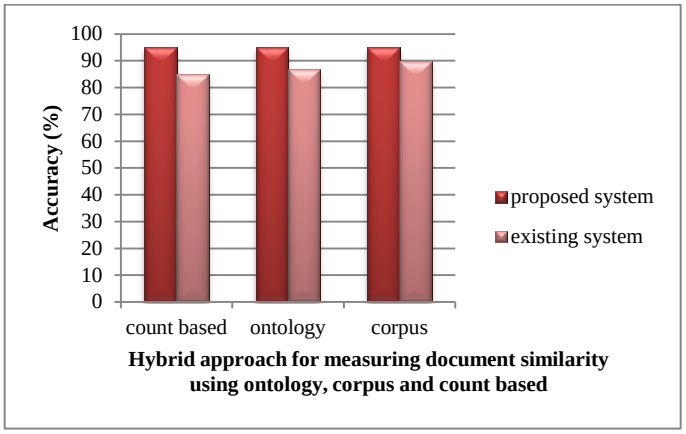

Fig. 5. Hybrid Approach for measuring document similarity using ontology, corpus and count based

The figure 5 shows Hybrid Approach for measuring document similarity using ontology, corpus and count based accuracy. When compare to existing and proposed the proposed approach has better accuracy.

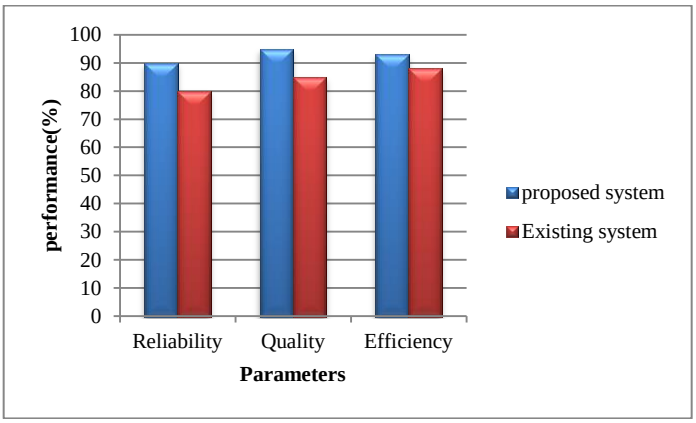

Fig. 6. Performance comparision

The figure 6 shows comparison of existing and proposed parameters paerformance. Here the parameters like reliability, quality and efficiency performance.

V. CONCLUSION

This work presented a hybrid approach for measuring semantic similarity between documents. Semantic similarity plays a crucial role in information retrieval and text processing. This work provided an overview of semantic similarity and its existing approaches. Semantic similarity measure determines the similarity between words, sentence, and documents. The proposed approaches are divided into two folds: In a first fold the proposed system exploits ontology-based similarity model and count based similarity model for measuring document similarity. In a second fold, the proposed system exploits ontology and corpus to estimate the document similarity. Due to the hybrid approach the proposed system achieves high accuracy in document similarity estimation.

REFERENCES

[1] D. Michie, "Return of the Imitation Game," Electronic Trans. Artificial Intelligence, Vol. 6, No. 2, pp. 203-221, 2001.

[2] E.K. Park, D.Y. Ra, and M.G. Jang, "Techniques for Improving Web Retrieval Effectiveness," Information Processing and Management, Vol. 41, No. 5, pp. 1207-1223, 2005.

[3] T.A.S. Coelho, P.P. Calado, L.V. Souza, B. Ribeiro-Neto, and R. Muntz, "Image Retrieval Using Multiple Evidence Ranking," IEEE Trans. Knowledge and Data Eng., Vol. 16, No. 4, pp. 408-417, 2004.

[4] J. Atkinson-Abutridy, C. Mellish, and S. Aitken, "Combining Information Extraction with Genetic Algorithms for Text Mining," IEEE Intelligent Systems, Vol. 19, No. 3, 2004.

[5] Moldovan, Dan I., and Rada Mihalcea. "A WordNet-Based Interface to Internet Search Engines", Proceedings of the Eleventh International FLAIRS Conference, pp. 275-279, 1998

[6] G. Erkan and D.R. Radev, "LexRank: Graph-Based Lexical Centrality As Salience in Text Summarization," J. Artificial Intelligence Research, Vol. 22, pp. 457-479, 2004.

[7] Y. Ko, J. Park, and J. Seo, "Improving Text Categorization Using the Importance of Sentences," Information Processing and Management, Vol. 40, pp. 65-79, 2004.

[8] Y. Liu and C.Q. Zong, "Example Based Chinese English MT," IEEE Proceedings IEEE Int'l Conf. Systems, Man, and Cybernetics, Vol. 1-7, pp. 6093-6096, 2004.

[9] Jacob B, Benjamin C(2008) "Calculating the Jaccard Similarity Coefficient with Map Reduce for Entity Pairs in Wikipedia", http://www.infosci.cornell.edu/weblab/papers/Bank2008.pdf

[10] Metzler, D., Bernstein, Y., Croft, W., Moffat, A., and Zobel, J. "Similarity measures for tracking information flow" Proceedings of CIKM, 517– 524, 2005.

[11] Dolan, W., Quirk, C., and Brockett, C. Unsupervised construction of large paraphrase corpora: Exploiting massively parallel new sources" In Proceedings of the 20th International Conference on Computational Linguistics, pp. 350- 356, 2004

[12] Allan, J., Bolivar, A., and Wade, C. "Retrieval and novelty detection at the sentence level" In Proceedings of SIGIR'03, 314–321, 2003.

[13] Hoad, T. and Zobel, J. "Methods for identifying versioned and plagiarized documents" Journal of the American Society of Information Science and Technology, Vol. 54, No. 3, pp. 203–215, 2003

[14] Chukfong H.,Masrah A. Azmi M.,Rabiah A. K, Shyamala C. Doraisamy. "Word sense disambiguation based sentence similarity", Proceedings of the 23rd International Conference on Computational Linguistics, pp. 418-426, 2003

[15] Tversky, A. "Features of similarity Psychological Review", Vol. 84, No. 4, pp. 327–352, 1977

[16] Shepard, R. N. and Arabie, P. "Additive clustering representations of similarities as combinations of discrete overlapping properties" Psychological Review, Vol. 86, No. 2, pp. 87–123, 1979

[17] Rohde, D. L. T. "Methods for binary multidimensional scaling" Neural Computation, Vol. 14, No. 5, pp. 1195– 1232, 2002.

[18] Lee, M. D. and Navarro, D. J. "An Empirical Evaluation of Models of Text Document Similarity" The annual conference of the Cognitive Science Society, pp. 1254- 1259, 2002.

[19] Yun .M, Jie L, and Zhengtao Y. "Concept Name Similarity Calculation Based on Wordnet and Ontology", Journal of software, Vol. 8, No.3, pp.746- 753, 2013.

[20] Rushdi S, and Adel E. "A Corpus-based Evaluation of a Domain-specific Text to Knowledge Mapping Prototype", Journal of computers, Vol. 5, No.1, pp. 69- 80, 2010.

[21] Deerwester, Scott C., et al., "Indexing by latent semantic analysis" Journal of the American Society for Information Science, Vol.41, No.6, pp. 391-407, 1990

[22] Royer, and Christiaan, "Term representation with generalized latent semantic analysis" proceedings in Recent Advances in Natural Language Processing, Vol.292, No. 45, 2005

[23] Gabrilovich, Evgeniy, and Shaul Markovitch. "Computing Semantic Relatedness Using Wikipedia-based Explicit Semantic Analysis", Proceedings of the 20th International Joint Conference on Artificial Intelligence, Vol. 7, pp. 1606-1611, 2007.

[24] Qin, P., Lu, Z., Yan, Y., et al., "A new measure of word semantic similarity based on WordNet hierarchy and DAG theory", International Conference on IEEE Web Information Systems and Mining, pp. 181-185, 2009.

[25] May Sabai Han, "Semantic Information Retrieval based on Wikipedia Taxonomy", International Journal of Computer Applications Technology and Research, Vol. 2, No.1, pp 77-80, 2013

[26] Sanchez, David, Montserrat Batet, et al., "Ontology-based semantic similarity: A new feature-based approach", Expert Systems with Applications, Vol. 39, No. 9, pp.7718- 7728, 2012.

[27] Zhang, Ce, et al., "Semantic similarity based on compact concept ontology" Proceedings of the 17th international conference on World Wide Web, ACM, pp. 1125-1126, 2008.

[28] Oleshchuk, Vladimir, and Asle Pedersen. "Ontology based semantic similarity comparison of documents", IEEE proceedings of 14th International Workshop on Database and Expert Systems Applications, 2003.

[29] Tous, Rubén, and Jaime Delgado, "A vector space model for semantic similarity calculation and OWL ontology alignment" In Database and Expert Systems Applications, Springer Berlin Heidelberg, Vol.4080, pp. 307-316, 2006.

[30] Peter D. Turney, "Measuring Semantic Similarity by Latent Relational Analysis", Proceedings of the 19th international joint conference on Artificial intelligence, pp. 1136-1141, 2005.

[31] Emiliano Giovannetti, Simone Marchi, and Simonetta Montemagni, "Combining Statistical Techniques and Lexico -syntactic Patterns for Semantic Relations Extraction from Text", SWAP 2008

[32] Liu, Gang, Ruili Wang, Jeremy Buckley, et al., "A WordNet-based Semantic Similarity Measure Enhanced by Internet-based Knowledge", In SEKE, pp. 175-178, 2011.

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