International Journal of Data Mining Techniques and Applications
Volume: 03 Issue: 02 December 2013, Page No.99-101
ISSN: 2278-2419

# Significance of Similarity Measures in Textual Knowledge Mining

K.L.Sumathy[1], S.Uma Shankari[2], Chidambaram[3]
[12]Research Scholar, BhararthyiarUniversity, Coimbatore
[3]Asst prof, Rajah serofiji College, Thanjavur.
Email:sumathyskcw@gmail.com

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

Keywords-similarity measures, structural similarity, semantic similarity, knowledge based similarity

## I. INTRODUCTION

Knowledge is scattered as explicit knowledge and unstructured knowledge, in the form of documents which are stored in the repositories. This form of knowledge is very difficult for the knowledge consumers. A system has to be evolved for knowledge discovery from text. This system should discover association, patterns of co-occurrence in a collection of textual documents. Text mining techniques can be classified in 2 ways 1. Inductive inference(text data mining),2. Abductive or Deductive inference(text knowledge mining) TKM would be of more useful in areas like Knowledge representation, reasoning algorithms for performing deductive and abductive inferences and knowledge based systems. In TKM the knowledge managed by the systems is collected from a collection of texts, each of one which is regarded as a particular knowledge base, these knowledge bases in texts are generated for the historical or normative reports. TKM provides new pieces of non-trivial ,unknown, and useful knowledge derived from a collection of text. TKM can be used for finding new knowledge without specifying a specific question or a query.

With the rapid growth of internet and the availability of text documents which paves way to automatically process documents for information extraction, interesting non-trivial information and to gather knowledge from unstructured text documents. Similarity measures have been used in Natural language processing and related areas. A document can be both structured and unstructured forms. Document similarity measures provides the capability to exploit the vast amount of unstructured information, such as documents and Web pages, in data repositories on Intranets and the Internet. It is estimated that unstructured information represents 80% of the total information available to a user. To address this "data digestion" problem, sophisticated solutions are required that turn unstructured data into information that knowledge workers can use to make informed decisions and can solve many problems related to documents. Relationship between knowledge mining and similarity measures is as follows.

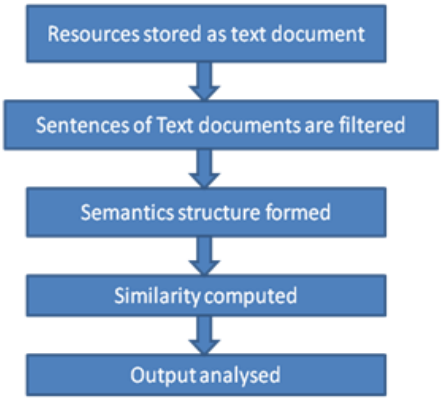


Figure-1

## II. VARIOUS TYPES OF DOCUMENT SIMILARITY MEASURES

There are various types of document similarity measures, namely, fulltext similarity, paragraph similarity, sentence similarity, semantic similarity, structural similarity, statistical similarity. Semantic similarity can be further classified into corpus based similarity and knowledge based similarity. Statistical similarity is further classified into jaccard, dice, cosine, levenshien edit distance, based on word distance and based on word vector. Structural similarity is classified into tree edit distance similarity, tag similarity, path similarity and Fourier transform similarity.

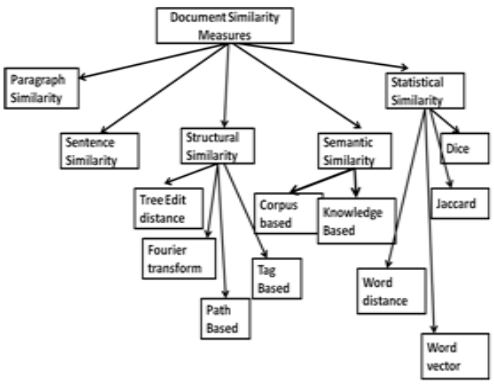


Figure-2

### A. SEMANTIC SIMILARITY

Measures of semantic similarity have been defined between words and concepts. One of the simple approach to find the

International Journal of Data Mining Techniques and Applications
Volume: 03 Issue: 02 December 2013, Page No.99-101
ISSN: 2278-2419

similarity between two segments of text is to use simple lexical matching method. Major contribution in this lexical matching is stemming, stop word removal, part-of-speech tagging. There are various measures of word to word semantic similarity measures, using approaches that are either knowledge or corpus based. For a text based semantic similarity measures, the most widely used approaches are through query expansion or thru latent semantic indexing methods. There are methods which target the identification of paraphrases starting with an input text.

Corpus-based Measures

Corpus-based measures of word semantic similarity identifies identify the degree of similarity between words using information from large corpora. There are 2 metrics namely(1) pointwise mutual information (Turney 2001), and (2) latent semantic analysis (Landauer, Foltz, &Laham 1998).

Pointwise Mutual Information.

The pointwise mutual information using data collected by information retrieval (PMI-IR) was suggested by (Turney 2001) as an unsupervised measure for the evaluation of the semantic similarityof words. It is based on word co-occurrence using counts collected on a large corpora. Given 2 words, w1 and w2, PMI-IR is calculated as follows

$\log_2$hit(w1 AND w2) * websize/(hits(w1) * hits(w2))

which indicates the degree of statistical dependence between w1 and w2, and can be used as a measure of the semantic similarity of w1 and w2.

Latent Semantic Analysis:

Latent Semantic Analysis is a corpus-based measure of semantic similarity is the latent semantic analysis (LSA) proposed by Landauer (1998). In LSA, term cooccurrences in a corpus are captured by means of a dimensionality reduction operated by a singular value decomposition (SVD) on the term-by-document matrix T representing the corpus. SVD is applied to any rectangular to find the correlations among its rows and columns.

Knowledge based semantic similarity measures.

There are various measures that were developed to quantify the degree to which two words are semantically related using information drawn from semantic networks .e.g. (Budanitsky&Hirst 2001) gives an overview. The Leacock &Chodorow (Leacock &Chodorow 1998) similarity is determined as:

$Sim_{KBSS}$=-log(length/2*D)

where length is the length of the shortest path between two concepts using node-counting, and D is the maximum depth of the taxonomy. The Wu and Palmer (Wu & Palmer 1994) similarity metric measures the depth of two given concepts in the Word Net Taxonomy and the least common subsume(LCS) and combines these figures into a similarity score.

$Sim_{wp}$=2*depth(LCS)/(depth(concept1)*depth(Concept2))the last similarity metric considered is Jiang &Conrath (Jiang &Conrath 1997)

$Sim_{JC}$= 1/(IC(concept1)*IC(Concept2)-2*IC(LCS))

## III. DOCUMENT STRUCTURE SIMILARITY ALGORITHMS

This section gives an overview of different types of algorithms used to determine the document similarity. The metrics described are tree edit distance similarity, tag similarity, Fourier transforms and path similarity.

Tree edit distance algorithm:

Several authors have proposed algorithms for computing the optimal edit distance between two trees. Optimal tree edit distance measures the minimum number of node insertions , deletions, and updates required to convert one tree into another. Let Ni be the set of nodes in the tree representation of document Di.

T ED(Di,Dj) = editDistance(Di,Dj)/max(|Ni|,|Nj|)

Tag similarity:

One of the simplest metric for structural similarity is tag similarity, as it measures only the set of tags match between two pages. The tag set of the two documents are compared to measure the overlap. Let Ti be the set of tags contained in page Dj. Tag similarity of the two pages is the intersection of the set of tags from Ti and Tj over the union. Let tik be a member of Ti, and wik be the number of times tag tik appears in Di. lettjk be the corresponding tag in Tj, and vjk be the number of times tag tjk appears in Dj . If there are n unique tags that occur in pages Di and Dj , then Weighted Tag

Similarity is calculated as

$$WTS(D_i, D_j) = \frac{\sum_{k=1}^{n} 2.min(w_{ik}, v_{ik})}{\sum_{k=1}^{n}(w_{ik}+v_{ik})}$$

Fourier transform similarity metric:

Flesca et al. introduced the Fourier transform technique as a mechanism to compute similarity between documents. The basic idea is to strip all the information from a document except for its start and end tags, leaving a skeleton that represents the structure. The structure is then converted into a sequence of numbers. The number sequence is then viewed as a time series, and a Fourier transform is applied to convert the data into a set of frequencies. Finally, the distance between two documents is computed by taking the difference of the magnitudes of the two signals. Flesca et al. chose a multilevel encoding of a document d as a sequence [S0;S1;:::;Sn] where

$$Si = \gamma(t_i) \text{x } expF(t_i) + \sum_{t_j \in next(t_i)} \gamma(t_j) \text{x } expF(t_j)$$

whereγ(ti) is the integer corresponding to the ith tag, expF (ti)= Bmaxdepth(D)..lti is an exponentiation factor determining the weight of the tag, where B is a fixed base, maxdepth(D) is the maximum depth of the documents being compared, lti is the depth of the ith tag, and nestd(ti) is the set of ancestors of ti. The final distance metric between two documents d1 and d2 using the Fourier transform is defined as

$$dist(d1,d2) = (\sum_{k=1}^{M/2}(|[FFT(h_1)](k) - |[FFT(h_2)](k)|)^2)^{\not\phi}$$

Path Shingles:

International Journal of Data Mining Techniques and Applications
Volume: 03 Issue: 02 December 2013, Page No.99-101
ISSN: 2278-2419

Shingles was introduced by Broder as a technique to compare two text documents for similarity.. The technique reduces the set of words, or tokens, in a document into a list of hashes that is directly compared with another document using set difference, union, and intersection to determine similarity or containment. A subset of the shingles, called a sketch, is used to determine document similarity. Sketches are a random sample of the text in a page. The key is that because the random mapping is constant across all pages, and the results are sorted, the samples are directly compared across different pages. The overlap in page samples indicates overlap between entire pages. Resemblance of a document is calculated using the formula

$$r(D_i, D_j) = \frac{S(D_i, w) \cap S(D_j, w)}{S(D_i, w) \cup S(D_j, w)}$$

containment is calculated using

$$c(D_i, D_j) = \frac{S(D_i, w) \cap S(D_j, w)}{S(D_i, w)}$$

1. **PROPOSED MODEL FOR DOCUMENT SIMILARITY:**

We have proposed an algorithm for detecting the similarity between 2 documents which is obtained by choosing a sentence in source document and compare it with the other sentences in the second document, if the similarity exceeds the specified range, it is defined as similar sentence, with the similar sentence, the proportion of document similarity is found.

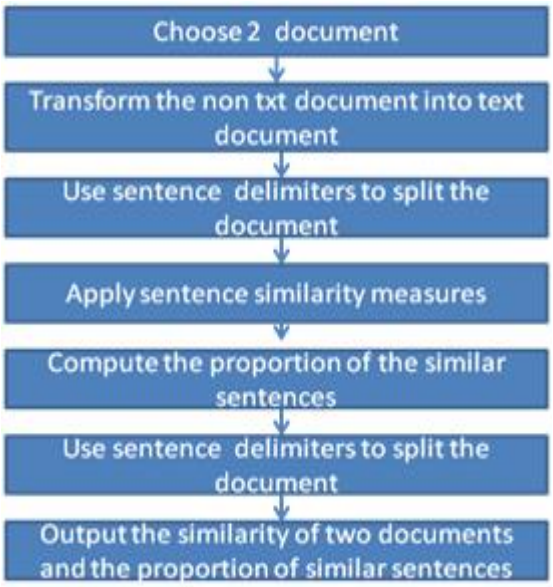


Figure-3

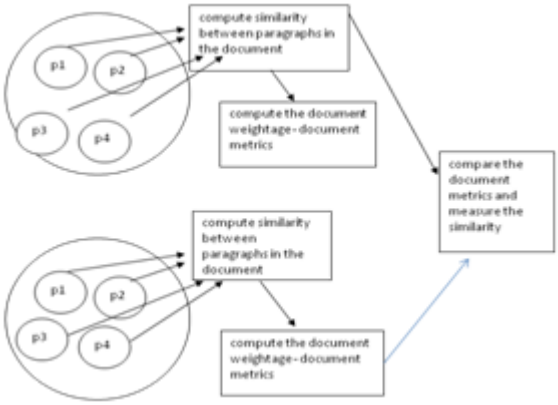


Figure-4

## IV. CONCLUSIONS

we have shown the different types of similarity for text documents based on the resemblance and containment. Document similarity is possible using a combination of semantic, layout and text features. These suggested various similarity measures for document can be used in the proposed model in implementation for further research.

## REFRENCES

[1] D. Buttler ,A Short Survey of Document Structure Similarity Algorithms March 5, 2004 The 5th International Conference on Internet Computing Las Vegas, NV, United States June 21, 2004 through June 24, 2004.
[2] Junsheng Zhang Calculating Statistical Similarity between sentences, Journal of Convergence Information Technology, Vol 6 No2 Feb 2011.
[3] Y. Wang, D. DeWitt, and J.-Y. Cai, .X-Diff: An effective change detection algorithm for XML documents," International Conference on Data Engineering, 2003.
[4] Barnard, C., and Callison-Burch, C. 2005. Paraphrasing with bilingual parallel corpora. In Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics.
[4]Barzilay, R., and Elhadad, N. 2003.Sentence alignment for monolingual comparable corpora. In Proceedings of the Conference on Empirical Methods in Natural Language Processing..
[5] Berry, M. 1992. Large-scale sparse singular value computations.International Journal of Supercomputer Applications 6(1).
[6]Budanitsky, A., and Hirst, G. 2001. Semantic distance in Word-Net: An experimental, application-oriented evaluation of five measures. In Proceedings of the NAACL Workshop on WordNet and Other Lexical Resources.
[7]Chklovski, T., and Pantel, P. 2004. Verbocean: Mining the Web for fine-grained semantic verb relations. In Proceedings of Conference on Empirical Methods in Natural Language Processing..
[8]Dagan, I.; Glickman, O.; and Magnini, B. 2005.The PASCAL recognising textual entailment challenge.In Proceedings of the PASCAL Workshop.
[9]Dolan, W.; Quirk, C.; and Brockett, C. 2004. Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In Proceedings of the 20th International Conference on Computational Linguistics.
[10]A. Marian, S. Abiteboul, G. Cobena, and L. Mignet, .Change centric management of versions in an XML warehouse," in The VLDB Journal, 2001, pp. 581.590.