IJRRECS/June 2014/Volume-2/Issue-6/2537-2542 **ISSN 2321-5461**

INTERNATIONAL JOURNAL OF REVIEWS ON RECENT ELECTRONICS AND COMPUTER SCIENCE

A SURVEY ON SUPPORTING SYSTEMS CONCERNING RETRIEVAL OF INFORMATION

E.Kondala Rao[1], M.S.R.Lakshmi Reddy[2]

[1]M.Tech Student, Dept of CSE, CMR Institute of Technology, Kandlakoya Medchal, Hyderabad, India

[2]Assistant Professor, Dept of CSE, CMR Institute of Technology, Kandlakoya Medchal, Hyderabad, India

ABSTRACT:

Identifying comparable pieces concerning text has numerous applications for instance summarization, retrieval of information and text clustering. Systems of information retrieval depend mostly on unambiguous, typed queries, collective with unambiguous feedback informing system which of search results was applicable. The criterion web search engines are basic versions of system; they take benefit of huge scale which permits inferring general interest concerning documents from link information. Measures of semantic resemblance have been conventionally defined among words otherwise concepts, and greatly less among text segments consisting of two or additional words. Present schemes of similarity measurement are extremely computationally demanding, making online scaling complicated.

Keywords: Text clustering, Semantic resemblance, Web Search engines, Search results.

1. INTRODUCTION:

The major general methods are latent semantic indexing as well as principal component analysis which analyze keywords concerning documents in a corpus to recognize the leading concepts in document [4]. we report effort on creating an online reverse dictionary in preference to a normal dictionary that map words in the direction of their definitions, a reverse dictionary as shown in fig1 carry out converse mapping, specifically specified a phrase describing needed concept, it makes available words whose definitions go with entered definition phrase [15]. We report the making of Wordster Reverse Dictionary, a

system of database-driven reverse dictionary which not merely fulfils novel efficient objectives at an order of extent performance and scale enhancement over finest notion resemblance measurement system obtainable devoid of impacting solution quality [9] [14]. Introduced reverse dictionary system is based on view that a phrase that describes a word have to bear a resemblance to the word's authentic definition, if not harmonizing the exact words, subsequently at least conceptually related [7] [12]. At a high level, introduced system consists of two sequential stages. Upon receiving of a client input phrase, initially discover candidate words against a forward source of dictionary data, where definitions concerning these candidate words encompass some resemblance towards user input [8]. Present schemes of semantic similarity measurement are extremely computationally demanding, making online scaling complicated [10]. We subsequently position candidate words in order of eminence of match. The phase concerning find candidate words phase comprises of two key sub steps such as building the RMS; and querying RMS [11] [13].

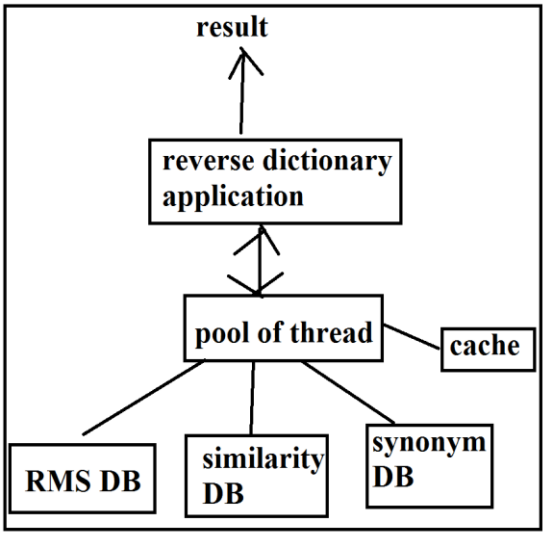

Fig1: An overview of reverse dictionary structure

II. LITERATURE SURVEY:

1. VanderMeer, and Kaushik Dutta [1] suggests that Reverse Dictionary Application is a software element that captures a user phrase as input, and returns theoretically connected words as output. This architecture has three features considered to make sure utmost scalability of system. A cache stock up regularly accessed information, which permits a thread towards accessing essential information devoid of contacting a database. It is renowned that several terms take place more commonly than others. The synonym, hyponym, hypernym, along with reverse map set sets of these well-liked terms will be stocked up in cache and query implementation in database will be

circumvented. The functioning of a thread pool permit in support of parallel recovery of synonym, hyponym, hypernym, as well as RMS sets in support of terms. Separate database augment opportunity in support of parallel processing, and augment system scalability. When a single machine is not competent of managing the essential loads, the database can effortlessly be additionally dispersed across numerous servers by means of partitioning methods to get better system scalability. To execute the processing the application of reverse dictionary requests accession towards information stored within a set of databases such as reverse map set DB, which enclose a table of mappings, in addition to dictionary definitions as well as computed parse trees in support of definitions; the Synonym DB, which enclose synonym set in support of each term; the Hyponym or Hypernym DB, containing hyponym as well as hypernym association sets for every term; the Antonym DB, enclose antonym set in support of every term; and authentic dictionary definitions in support of every word in dictionary. The mappings in support of reverse map set, synonyms, hyponyms, hypernyms, as well as antonyms are stored up as integer mappings, where every word in Wordnet dictionary is symbolized by means of a exceptional integer. This condenses size of the mapping sets, and allow in support of extremely fast processing of resemblance comparisons, as evaluated to string processing. When reverse dictionary application desires word-related statistics, it assign the job of recovering this data towards a thread within a thread pool. The thread initially monitors local cache to conclude whether suitable data subsist in cache. If so, cache returns essential information. If not, cache returns a null set. When threads receive a null set concerning cache, it contacts suitable database to get hold of the essential information.

2. J. Klavans, and E. Eskin [2] suggests that identifying comparable pieces concerning text has numerous applications for instance summarization, retrieval of information and text clustering. For the most part of research in has centered on sensing resemblance among documents, resemblance connecting a query with a document or connecting a query and a segment concerning a document. While effectual techniques have been expanded in support of document clustering as well as classification which rely on inter-document

resemblance measures, these methods mainly rely on pooled words, or intermittently collocations concerning words. When outsized units concerning text are evaluated, overlap might be enough to become aware of resemblance; but when the units concerning text are minute, effortless surface matching of words as well as phrases is less probable to make it while number of possible matches is less significant.

3. David R. Hardoon, John Shawe-Taylor [3] put forward that present systems of information retrieval depend mostly on unambiguous, typed queries, collective with unambiguous feedback informing system which of search results was applicable. The relevance response is used to improve query, and search congregate iteratively towards additionally pertinent documents. The criterion web search engines are basic versions of system; they take benefit of huge scale which permits inferring general interest concerning documents from link information. Usage of eye movements within information retrieval is a reasonably novel approach. Maglio and Campbell set up a sample attentive agent function which observes eye movements although user views web pages, to settle on whether the user understands. When reading is noticed, additional information concerning topic is displayed. The possibility of application was not on the other hand experimentally confirmed.

4. D. Lin [5] suggests that feature vectors are simplest and most generally used forms of knowledge depiction, particularly in case based reasoning as well as machine learning. Weights are regularly allotted to features to account for information that difference caused by additional significant features is superior to difference caused by less significant features. The mission of the weight parameters is commonly heuristic in character in preceding approaches. Resemblance is a basic and extensively used notion. Numerous resemblance measures have been projected, for instance information content, mutual information, cosine coefficient, distance-based measurements as well as feature contrast representation. A trouble with preceding resemblance measures is that each of them is coupled to a meticulous application or believes a meticulous domain representation. If an assortment of documents is not symbolized as a network,

distance-based process does not be valid. The Dice as well as cosine coefficients are appropriate merely when objects are characterized as numerical feature vectors.

5. C. Corley, and C. Strapparava [6] recommends that text resemblance has been used in support of relevance feedback as well as text classification, word sense disambiguation and more lately in support of extractive summarization and methods in support of automatic assessment of machine translation otherwise text summarization. Measures concerning text resemblance were also useful for assessment of text coherence. Measures of text resemblance were used for a long time in application in usual language processing and connected areas. One of most basic applications concerning text similarity is possibly vectorial representation in information recovery; where document most pertinent to an input query is indomitable by means of ranking documents in an assortment in inverted order of their resemblance to specified query. Measures of semantic resemblance have been conventionally defined among words otherwise concepts, and greatly less among text segments consisting of two or additional words. The importance on word-to-word resemblance metrics is probably due to accessibility of resources that particularly encode relations connecting word otherwise concepts, and the variety of test beds that permit for their assessment. The derivation concerning a text-to-text assess of similarity starting with a word based semantic resemblance metric may not be simple, and thus for the most part of the work has measured mainly function of conventional vectorial representation.

III. CONCLUSION:

Text resemblance has been used in support of relevance feedback as well as text classification, word sense disambiguation and more lately in support of extractive summarization and methods in support of automatic assessment of machine translation otherwise text summarization. Measures of text resemblance were used for a long time in application in usual language processing and connected areas. The two major general methods are latent semantic indexing as well as principal component analysis which analyze keywords concerning documents in a corpus to recognize the leading concepts in document. Reverse Dictionary Application is a software element that captures a user phrase as input, and returns theoretically

connected words as output. It carry out converse mapping, specifically specified a phrase describing needed concept, it makes available words whose definitions go with entered definition phrase.

REFERENCES:

[1] "Building a Scalable Database-DrivenReverse Dictionary", Ryan Shaw,, Anindya Datta, Debra VanderMeer, and Kaushik Dutta, 2013

[2] V. Hatzivassiloglou, J. Klavans, and E. Eskin, "Detecting Text Similarity over Short Passages: Exploring Linguistic FeatureCombinations Via Machine Learning," Proc. Joint SIGDAT Conf.Empirical Methods in Natural Language Processing and Very LargeCorpora, pp. 203-212, June 1999.

[3] "Information Retrieval by Inferring Implicit Queries from Eye Movements", David R. Hardoon, John Shawe-Taylor, Antti Ajanki, Kai Puolam¨aki, Samuel Kaski, 2007.

[4] T. Dao and T. Simpson, "Measuring Similarity between Sentences," http://opensvn.csie.org/WordNetDotNet/trunk/Projects/Thanh/Paper/WordNetDotNet_Semantic_Similarity.pdf(last accessed 16 Oct. 2009), 2009.

[5] D. Lin, "An Information-Theoretic Definition of Similarity," Proc.Int'l Conf. Machine Learning, 1998.

[6] R. Mihalcea, C. Corley, and C. Strapparava, "Corpus-Based andKnowledge-Based Measures of Text Semantic Similarity," Proc.Nat'l Conf. Artificial Intelligence, 2006.

[7] R. Nallapati, W. Cohen, and J. Lafferty, "Parallelized Variationalem for Latent Dirichlet Allocation: An Experimental Evaluation ofSpeed and Scalability," Proc. IEEE Seventh Int'l Conf. Data MiningWorkshops, pp. 349-354, 2007.

[8] P. Resnik, "Semantic Similarity in a Taxonomy: An Information-Based Measure and Its Application to Problems of Ambiguity inNatural Language," J. Artificial Intelligence Research, vol. 11, pp. 95-130, 1999.

[9] J. Carlberger, H. Dalianis, M. Hassel, and O. Knutsson, "ImprovingPrecision in Information Retrieval for Swedish Using Stemming,"Technical Report IPLab-194, TRITA-NA-P0116, Interaction and Presentation Laboratory, Royal Inst. of Technology andStockholm Univ., Aug. 2001.

[10] J. Lafferty and C. Zhai, "Document Language Models, QueryModels, and Risk Minimization for Information Retrieval," Proc.24th Ann. Int'l ACM SIGIR Conf. Research and Development inInformation Retrieval, pp. 111-119, 2001.

[11] G. Salton, J. Allan, and C. Buckley, "Approaches to PassageRetrieval in Full text Information Systems," Proc. 16th Ann. Int'lACM SIGIR Conf. Research and Development in Information Retrieval,pp. 49-58, 1993.

[12] J. Kim and K. Candan, "Cp/cv: Concept Similarity Miningwithout Frequency Information from Domain Describing Taxonomies,"Proc. ACM Conf. Information and Knowledge Management,
2006.

[13] J. Ponte and W. Croft, "A Language Modeling Approach toInformation Retrieval," Proc. 21st Ann. Int'l ACM SIGIR Conf.Research and Development in Information Retrieval, pp. 275-281, 1998.

[14] T. Hofmann, "Probabilistic Latent Semantic Indexing," SIGIR '99:Proc. 22nd Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval, pp. 50-57, 1999.

[15] H. Cui, R. Sun, K. Li, M.-Y. Kan, and T.-S. Chua, "QuestionAnswering Passage Retrieval Using Dependency Relations," Proc.28th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval, pp. 400-407, 2005.