ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 10, October 2014

A Web Search Engine-Based Approach to Measure Semantic Similarity between Words

Narendra Pradhan[1], Kamlesh kumar Pandey[2], Rajesh sahu[3]

Assistant Professor, Computer Science, S.S.College of Education, Pendra Road, India[1]

Assistant Professor, Computer Science, Makhanlal University Amarkantak, India[2]

Assistant Professor, Computer Science, S.S.College of Education, Pendra Road, India[3]

Abstract:A web search engine is software code that is designed to search for information on the World Wide Web. The search results are generally presented in a line of results often referred to as search engine results pages (SERP's). The information may be a specialist in web pages, images, information and other types of files. Some search engines also mine data available in data basesor open directories. Semantic similarity or semantic relatedness is a concept whereby a set of documents or terms within term lists are assigned a metric based on the likeness of their meaning / semantic content. The existing technique implemented for the sematic similarity of the words is based on novel pattern extraction algorithm and a pattern clustering algorithm. Here we are implementing an efficient algorithm for the searching of the similarity of the words using Machine Learning based rule approach.

Keywords: Web, Search Engine, Measure, Semantic.

1. INTRODUCTION

The study of semantic similarity between words has been a part of natural language processing and information retrieval.Accurately measure the linguistics similarity between words is a very importantdownside in internet mining, data retrieval, and languageprocess. Internet mining applications like community extraction, relation detection, and entity disambiguation; needthe power to accurately live the linguistics similarity between ideas or entities. In data retrieval, one in every ofthe mostissues is to retrieve a collection of documents that's semantically associated with a given user question. Economical estimation of linguistics similarity between words is vital for numerouslanguageprocess tasks likeacceptationelucidation (WSD), matterinference, and automatic text summarisation.

A search engine operates in the following order:

1. Web crawling
2. Indexing
3. Searching

Web search engines work by storing infoconcerningseveralwebsites that they retrieve from the hypertext mark-up language itself. These pages area unit retrieved by an online crawler (sometimes additionallyreferred to as a spider) — an automaticapplications programmethat follows each link on the location. Exclusions are oftencreated by the utilization of robots.txt. The contents of every page area unit then analysedto seehowever it ought to be indexed (for example, words are often extracted from the titles, page content, headings, or special fields referred to as meta tags).knowledgeconcerningwebsitesarea unitkeep in associate index informationto be used in later queries. aquestion are oftenone word. The index helps noticeinfo as quickly as attainable. Some search engines, like Google, store all or a part of the supply page (referred to as a cache) likewise as infoconcerningthe online pages, whereas others, like AltaVista, store each word of each page they notice. This cached page forever holds the particular search text since it's the one that was truly indexed, thus it are oftenterriblyhelpfulonce the content of this page has been updated and therefore the search terms aren't any longer in it. This downsidecan bethought ofa gentlesort of link rot, and Google's handling of it will increase usability by satisfying user expectations that the search terms are on the came back webpage. This satisfies the principle of least feeling, since the user unremarkably expects that the search terms are on the came back pages. Increased search relevance makes these cached pages very useful, even beyond the fact that they may contain data that may no longer be available elsewhere.

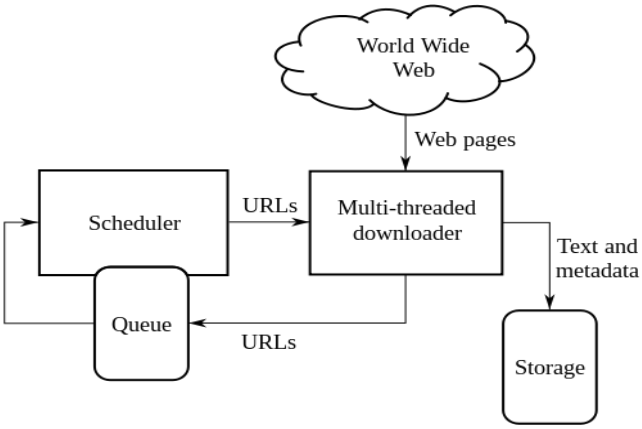

Figure 1. Architecture of Web

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 10, October 2014

Semantically related words of a particular word are listed in manually created general-purpose lexical ontologies such as WordNet.1 In WorldNet, a synset contains a set of synonymous words for a particular sense of a word. However, semantic similarity between entities changes over time and across domains. For example, apple is frequently associated with computers on the web. However, this sense of apple is not listed in most general-purpose thesauri or dictionaries. A user who searches for apple on the web, might be interested in this sense of apple and not apple as a fruit. New words are constantly being created as well as new senses are assigned to existing words. Manually maintaining ontologies to capture these new words and senses is costly if not impossible [1].

The study of linguistics similarity between words has been a section of linguistic communicationprocessand knowledge retrieval for several years. Linguistics similarity [2] could be a generic issue in an exceedinglystyle of applications within the areas of linguistics and computer science, eachwithin thetutorial community and business. Examples embodysenseclarification, detection and correction of word writing system errors (malapropisms) [3], text segmentation, image retrieval, multimodal documents retrieval, and automatic machine-readable text linking. Similarity between 2 words is commonlyportrayed by similarity between ideasrelated tothe 2 words. Variety of linguistics similarity waysis developed within the previous decade; totally different similarity ways have verified to be helpful in some specific applications of process intelligence. Generally, these waysmay becategorised into 2 groups: edge counting- primarily based (or dictionary/thesaurus based) waysand knowledgetheoretical (or corpus based) ways [4].

The information theory-based method for semantic similarity was first proposed by Resnik [5]. He defines the similarity of two concepts as the maximum of the information content of the concept that subsumes them in the taxonomy hierarchy. The information content of a concept depends on the probability of encountering an instance of the concept in a corpus. That is, the probability of a concept is determined by the frequency of occurrence of the concept and its sub concept in the corpus. The information content is then defined as negative the log likelihood of the probability. Since the information content is calculated from the corpus, this similarity measure can be adapted to a particular application provided that the corpus approximates that application area well.

Numerous info retrieval and linguistic communicationprocess applications needinformation of linguistics similarity between words or terms. As an example, by adding semantically similar words to an internetquestion (query expansion), it'sprobablyto extend the relevance1 of retrieved documents. Moreover, linguistics similarity measures areemployed inseverallinguistic communicationprocess (NLP) tasks, like language modelling , synchronic linguistics induction, senseelucidation , and speech understanding and spoken dialogue systems .several unsupervised applied mathematics metrics arebestowed and applied to the automated induction of linguisticscategories for each semantically consistent and heterogeneous corpora [7].

The Web has a multilingual character: new words, neologisms, and occasionalisms (hap ax ligament) are added frequently and efficiently. Thus, it is the obvious place for mining semantic relationships for unseen words. The Web also contains both general purpose words, found in news articles and blogs, as well as scientific terminology, found in documents written by experts. Overall, the Web covers a plethora of domains, authoring styles and languages, and is fertile ground for automatic semantic knowledge acquisition. The Web has been exploited for a variety of NLP applications. Webpage counts returned by a search engine were used to estimate the probability of n-gram language models. The Webpage counts of fixed lexical patterns were used to identify synonymy and antonym between nouns. Web queries of lexicon-syntactic patterns were used for discovering relationships between verbs. The Web is also an invaluable source for constructing text corpora. A large corpus of Web Pages was constructed and used for word sense disambiguation. Other applications, where automatically constructed Web corpora have been used to train statistical models, include machine translation [8] and question-answering systems.

2. Semantic Similarities between Words

Before continuing to the presentation of our methodology, it's necessary to introduce some constraints to the event of similarity measures. evidence from psychological experiments demonstrates that similarity is context-dependent and will be uneven. Similarity between words is influenced by the context during which the words aregiven. As an example, if the context is “the outside covering of living objects,” then skin and bark area lot of similar than skin and hair; but, the other is true if the context is body elements. Similarity may additionally be asymmetric with relevancy direction. Peoplecouldprovidecompletely different ratings once asked to gauge the similarity of surgeon to butcher and also the similarity of butcher to surgeon. Though similarity is alsouneven, the “asymmetries aresolelydeterminedbeneath quite circumscribed conditions”. Experimental results workthe consequences of imbalancerecommend that the commondistinction in ratings for a word combineis a smaller amount than five%. we tend to believe that such a littledistinctioncan have very little impact on the performance of procedurestrategies, thuswe tend todon'tthink aboutthe consequences of imbalance. this can be in line with several application areas of linguistics and computing [2].

2.1 The Method for Semantic Similarity

. The data bases could also becreatedin an exceedingly hierarchy that's commonplace within the

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 10, October 2014

world. The lexical hierarchy is connected by following trails of super ordinate terms in "is a" or "is a form of" (ISA) relations. The ISA hierarchical data structure of the knowledge domainis vital in determinative the linguistics distance between words. Fig. two shows some of such a stratifiedlinguisticsknowledge domain [2].

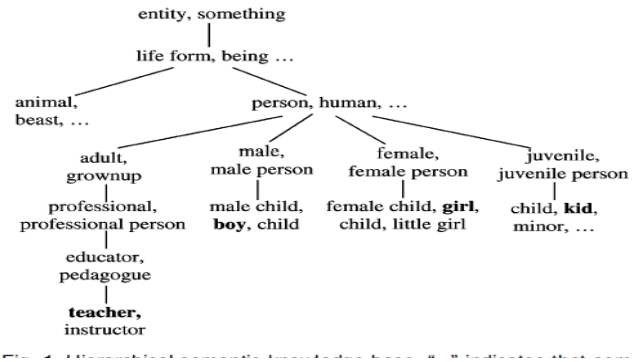

Fig. 2. Hierarchical semantic knowledge base. "..." indicates that some words in the class were omitted to save space.

RELATED WORK

In 2011 byDanushka Bollegala et. Al.proposean practical method to approximation semantic similarity using page counts and text snippets retrieved from a web search engine for two words. Specifically and define various word co-occurrence measures using page counts and integrate those with lexical patterns extracted from text scraps. To identify the numerous semantic relations that exists between two given words, and proposes a novel pattern extraction algorithm and a pattern of clustering algorithm. The best combination of page counts-based co-occurrence procedures and lexical pattern clusters is academic using support vector equipment. The proposed method performs various baselines and web-based semantic similarity procedures on three benchmark data sets showing a high correlation with person ratings. besides, the proposed method significantly improves the accuracy in a community mining task [1].

In 1999 by the authors David Castanon et al. prove that the optimal feedback strategies for this problem are index policies and provide an explicit expression for the best expected reward from any situation. The problem is forced by search methods for global optimization problems where the cost of computation is explicitly incorporated into the objective. The objective of the problem is to maximize the expected net difference between the largest sample reward obtained before stopping and the accumulated costs incur while sampling [9].

Author introducing the user controlled classification methods in addition to text search and filtering for increasing recollect in analytics scenarios involving large corpora. Classification through machine learning has the prospective to improve search and filter tasks around either complex or very specific information needs, individually. Large amount classification methods, on the other hand, require a assured expertise concerning their parameterization to achieve high-quality results. Supervised machine learning algorithms, in difference, rely on labeled data, which can be provided by analysts and presented an approach for building classifiers interactively, and visually in order to complement classical search and filter techniques, which often show weaknesses regarding recall and generalization. here main contribution in suggesting the described approach as a whole, starting from bootstrapping an initial classifier using a keyword query, to classifier training employing activelearning methods, and its future re-integration into keyword queries. The design of the described evaluation further fuels the discussion on employing baselines for comparative test procedures for visual interactive classifier creation [10].

Inside these concept using different methods they are: 1: The Basic Method, 2: The Visual Method, 3: The User-Driven Method.

1: The Basic Method

The Basic Method realizes textual AL without employing sophisticated visualization. Analysts have to label the text presented to them as relevant or non-relevant according to the given information need. The tool (see Fig. 3) provides a Search Bar, where analysts can enter an initial query and start the classifier building procedure with the button next to it. The headline following the text field has been specifically introduced for our user evaluation and informs the participants about the current search task.

2: The Visual Method

The Visual Method provides users with feedback on the classifier's state and lets them explore the classification context by representing the whole set of documents including labeled and unlabeled ones. The Visual Method's tool features multiple linked views that provide analysts with insights on the data and the classifier's state. The Visual Method's tool contains the Search Bar for entering the initial query and a button to generate the initial classifier (see Fig. 1a). The Main View (Fig. 1b) shows a graphical representation of the current state of the classifier in form of a scatter plot.

3: The User-Driven Method

The User-Driven Method no longer uses AL directly, but incorporates some ideas derived from it. Here, analysts have full control over the selection of the documents they want to label. It also allows the labeling of multiple documents at once. The Main View and the Cluster View now provide interaction mechanisms for selecting single documents, by clicking on them, and multiple documents with a rectangular selection mechanism [10].

Here author describes a web mining method to classify research documents mechanically. Web hit counts

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 10, October 2014

of AND search consist of two words are used to form a text vector. Target ID are classified with a result of k-means clustering method, in which cosine connection is used to calculate a distance. It uses AND-search on two words from representative words of each document. And it utilizes both tiff value of each word, the cosine similarity and k-means clustering method. The algorithm is simple. But our preliminary experimental result shows that it produces a classification, which is not random. One of key points to improve this algorithm is the selection of representative words. At this point, it uses only a simple filtering of words. Some other criteria to filter words are required [11].

Inside this concept using the different types of method include:

TFIDF Weight Factor

The tiff is stand for frequency-inverse document frequency, where there is a weight factor to show how main a word is to a document in a document set or in a dictionary. The tf is a frequency value, which is the number of time a word occurs in a document divided by the total number of words in the document. The ide is an inverse of a frequency value, which is the number of documents where the word is included divided by the total number of documents. Tiff value is calculated by multiplying tf value and idf value. After calculating thief weight factor for each word in a document, our algorithm removes some kind of words. It is an ad hoc choice at this point. At first, it removes short words less than three characters. They seems to be just names of variables.

AND-search on the web:

All representative words from each document are mixed. And web hit counts of AND-search with any two words from the mixed words are retrieved and a table of hit counts is created.

K-means Clustering:

Our method uses the k-means clustering for assigning of representative words. It is a method of cluster analysis, which is based on calculation of Euclidean distance. Given n data points and the number of clusters k, this algorithm allocates the n data points to k clusters [11]. This algorithm begins by selecting an initial set of k data points. Each of the set becomes a reference point of a cluster. And then it repeats the following two steps.

• For every n data points, compute the adjacent reference point and allocate it to the cluster related to the nearest.
• For every new cluster, analyze the mean to be the new indication point in the cluster.

In 2012 by Raja Sunkara et al. gives the concept about sentence similarity words from the sentences are considered and their respective taxonomies are built with Word Net. The evolved taxonomies are combined to develop Hierarchical Ontology. The comparison is done with the help of an empirical formula (SenSim) on the Hierarchical Ontology developed from the two sentences and found that our proposed method gives fairly good sentence similarity measure. Sentence similarity measure is an important concern for researchers for text retrieval in areas just like text mining, web information recovery, decision making and question identical. Accessible methods for check out sentence similarity have been adopted. To find out the semantic similarity calculate, the gap between two concepts, wisdom level, neighbor nodes and sibling factor of the ontology are considered in framing the formula. We provide an effective approach of finding semantic similarity between simple sentences by considering the problem statement of three sentences, to reduce complexity. This work can be further extended with more number of sentences in the domain set to find out the similarity measure. The accuracy and effectiveness of the model is reflected by the computational results in identifying the more or less similar sentences dependent upon the threshold values [12].

By Nattakarn Ratprasartporn et al. proposes a replacement literature digital assortment search model that effectively ranks search outputs, whereasdominantthe range of keyword-based search question output topics. Here approach is as follows. First, throughout pre-querying, publications arappointed into pre-specified ontology-based contexts, and query-independent context scores arnear papers with reference to the appointed contexts. Oncea question is expose, relevant contexts arelite, search is performed amongthe chosen contexts, context legion publications ar revised into connectedness scores with reference to the question at hand and therefore the context that they're in, and question outputs arstratifiedamongevery relevant context. This way, we tend to (1) minimize question output topic diversity, (2) scale backquestion output size, (3) decrease user time spent scanning question results, and (4) increase question output ranking accuracy. victimisation genomics-oriented Pub Med publications because the tested and factormetaphysics terms as contexts, our experiments indicate that the projected context-based search approach produces search results with up to five hundredth higher exactness, and reduces the question output size by up to seventieth [13].

In one contextual web search approach, a context is captured around the user-highlighted content, and amplified queries are produced from the selected context words. This approach is similar to our context-based search approach in the sense that users can specify contexts of interests before viewing search results. The main differences are that the contexts of this approach come from documents as opposed to a pre-defined ontology-based hierarchy, and no structural and hierarchical information are used [14].

Another technique, called Tile Bars [15], lets the user enter a query in a faceted format (i.e., each line represents each topic) and provides graphical bar in order to show the degree of match for each facet. TileBars illustrate which parts of each document contain which topic by dividing the bar into columns, where each column refers to a part in the document. The darkness of the square indicates the number of times the topic occurs in

the part of the document. With this approach, the user can easily see the relevancy of the document to each specified topics. On the other hand, search results are exposed as one list and no categorization of search results is provided.

A number of categorization techniques have been proposed to make search results more understandable. Two widely-used categorization techniques are document clustering and document classification. Document clustering creates categories (or contexts) by grouping similar documents together while document classification assigns documents to a set of predefined categories [16].

REFERENCES

[1]. Danushka Bollegala, Yutaka Matsuo, And Mitsuru Ishizuka'" A Web Search Engine-Based Approach To Measure Semantic Similarity Between Words", Ieee Transactions On Knowledge And Data Engineering, Vol. 23, No. 7, July 2011.

[2]. Yuhua Li, Zuhair A. Bandar, And David Mclean," An Approach For Measuring Semantic Similarity Between Words Using Multiple Information Sources" IEEE Transactions On Knowledge And Data Engineering, Vol. 15, No. 4, July/August 2003.

[3]. A. Budanitsky and G. Hirst, "Semantic Distance in Word Net: An Experimental, Application-Oriented Evaluation of Five Measures," Proc. Workshop Word Net and Other Lexical Resources, Second Meeting North Am. Chapter Assoc. for Computational Linguistics, June 2001

[4]. A.W.M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-Based Image Retrieval at the End of the Early Years," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 22, no. 12, pp. 1349-1380, Dec. 2000.

[5]. P. Resnik, "Using Information Content to Evaluate Semantic Similarity in a Taxonomy," Proc. 14th Int'l Joint Conf. Artificial Intelligence, 1995.

[6]. Elias Iosif, Alexandros Potamianos," Unsupervised Semantic Similarity Computation Between Terms Using Web Documents", Ieee Transactions On Knowledge And Data Engineering, Vol. 22, No. 11, November 2010.

[7]. E. Iosif, A. Tegos, A. Pangos, E. Fosler-Lussier, and A. Potamianos, "Combining Statistical Similarity Measures for Automatic Induction of Semantic Classes," Proc. IEEE/ACL Workshop Spoken Language Technology, pp. 86-89, 2006.

[8]. M. Popovic and H. Ney, "Exploiting Phrasal Lexica and Additional Morpho-Syntactic Language Resources for Statistical Machine Translation with Scarce Training Data," Proc. 10th Ann. Conf. European Assoc. for Machine Translation, pp. 212-218, 2005.

[9].David Casta˜Non, Simon Streltsov, And Pirooz Vakili," Optimality Of Index Policies For A Sequential Sampling Problem", Ieee Transactions On Automatic Control, Vol. 44, No. 1, January 1999.

[10].Florian Heimerl, Steffen Koch, Harald Bosch,," Visual Classifier Training For Text Document Retrieval", Ieee Transactions On Visualization And Computer Graphics, Vol. 18, No. 12, December 2012.

[11]. Masaya Kaneko, Shusuke Okamoto, Masaki Kohana "Document Classification based on Web Search Hit Counts" iiWAS2012 3-5 December, 2012, Bali, Indonesia Copyright 2012 ACM 978-1-4503-1306.

BIOGRAPHY

Narendra Pradhan, Completed MSc-IT From GGCU Bilaspur C.G. I am Presently Working as Assistant Professor in Department of Computer Science, S.S.College Of Education Sarbahra Pendra Road Bilaspur. I am Having 2 years of teaching experience My interested Subject are Computer Network, Network Security, Operating System and Web technology.