International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2015): 6.391

Information Retrieval Using Semantic Distance between WordNet

Rahul Shirbhate[1], Vishal Mogal[2]

[1]M.E. Student, Dept. of Computer, RMD Sinhgad School of Engineering, Pune, India

[2]Assistant Professor, Dept. of Computer, RMD Sinhgad School of Engineering, Pune, India

Abstract: ***Semantic search seeks to improve search accuracy by understanding searcher intent and the contextual meaning of terms as they appear in the searchable data space, whether on the Web or within a closed system, to generate more relevant results. There are 11 approaches that join semantics to search and provide an overview that lists semantic search systems and identifies other uses of semantics in the search process. Semantic search systems consider various points including context of search, location, intent, and variation of words, synonyms, generalized and specialized queries, concept matching and natural language queries to provide relevant search results. Major web search engines like Google and Bing incorporate some elements of semantic search. The semantic web is a idea of teaching the web which is to say teaching the next generation of web search engines and web browsers how to understand the content rather than just the structure on the web.***

Keywords: geographic concepts; semantic similarity; geo-knowledge graphs; network-lexical similarity measure (NLS); lexical similarity; network similarity

1. Introduction

A Web Search Engine is designed to search for information on the World Wide Web and FTP servers. The search results are generally presented in a list of results and are often called hits. The information may consist of web pages, images, information and other types of files. Some search engines also mine data available in databases or open directories. A program that searches documents for specified keywords and returns a list of the documents where the keywords were found. Although search engine is really a general class of programs, the term is often used to specifically describe systems like Google, Alta Vista and Excite that enable users to search for documents on the World Wide Web and USENET newsgroups. Typically, a search engine works by sending out a spider to fetch as many documents as possible. Another program, called an indexer, then reads these documents and creates an index based on the words contained in each document. Each search engine uses a proprietary algorithm to create its indices such that, ideally, only meaningful results are returned for each query.

2. Literature Survey

All the paper presents an innovative architecture for semantic search engine. Five different proposed measures of similarity or semantic distance in WordNet were experimentally compared by examining their performance in a real-world spelling correction system.

Alexander Budanitsky, Graeme Hirst," *Semantic distance in WordNet: An experimental, publication-oriented evaluation of five measures*". It was found that Jiang and Conraths measure gave the best results overall. That of Hirst and St-Onge seriously over-related, that of Resnik seriously under-related, and those of Lin and of Leacock and Chodorow fell in between. The need to determine the degree of semantic similarity, or, more generally, relatedness, between two lexically expressed concepts is a problem that pervades much of computational linguistics. Measures of similarity or relatedness are used in such applications as word sense disambiguation, determining discourse structure, text summarization and annotation, information extraction and retrieval, automatic indexing, lexical selection, and automatic correction of word errors in text. The problem of formalizing and quantifying the intuitive notion of similarity has a long history in philosophy, psychology, and artificial intelligence, and many different perspectives have been suggested.

Recent research on the topic in computational linguistics has emphasized the perspective of semantic relatedness of two lexemes in a lexical resource, or its inverse, semantic distance. It's important to note that semantic relatedness is a more general concept than similarity; similar entities are usually assumed to be related by virtue of their likeness (bank-trust company), but dissimilar entities may also be semantically related by lexical relationships such as metonymy (car-wheel) and antonym (hot-cold), or just by any kind of functional relationship or frequent association (pencil-paper, penguin-Antarctica). Computational applications typically require relatedness rather than just similarity; for example, money and river are cues to the in-context meaning of bank that are just as good as trust company. However, it is frequently unclear how to assess the relative and absolute merits of the many competing approaches that have been proposed. Our purpose in this paper is to compare the performance of several measures of semantic relatedness that have been proposed for use in NLP applications.

Daniel Fried, Kevin Duh," *Incorporating Both Distributional And Relational Semantics In Word Representations*". It describes, the ideal search engine would be able to match the search queries to the exact context and return results within that context. While Google, Yahoo and

Live continue to hold sway in search, here are the engines that take a semantics (meaning) based approach, the end result being more relevant search results which are based on the semantics and meaning of the query Semantic search seeks to improve search accuracy by understanding searcher intent and the contextual meaning of terms. Rather than using ranking algorithms such as Google's PageRank to predict relevancy, Semantic Search uses semantics, or the science of meaning in language, to produce highly relevant search results. In most cases, the goal is to deliver the information queried by a user rather than have a user sort through a list of loosely related keyword results.

Abdeslem D., Sidi Mohammed, "*A New Measure of the Calculation of Semantic Distance between Ontology Concepts*", It describes There is a substantial trend in the evolution of online search as people become more sophisticated in their knowledge of search engines and learn that more specific search terms (usually longer search phrases) deliver more accurate search results and in simple search user does not get that record or data which he want. Also the semantic web is to make the meaning of information explicit through semantic mark-up, thus enabling more effective access to knowledge contained in heterogeneous information environments, such as the web. Semantic search plays an important role in realizing this goal, as it promises to produce precise answers to user's queries by taking advantage of the availability of explicit semantics of information. For example, when searching for news stories about Phd students, with traditional searching technologies, we often could only get news entries in which the term Phd students appears.

Jaap Kamps, Maarten Marx, Robert J. Mokken, Maarten de Rijke, "*Using WordNet to Measure Semantic Orientations of Adjectives*", The semantic web is the idea of teaching the web which is to say teaching the next generation of web search engines and web browsers how to understand the content rather than just the structure on the web. It is teaching the engines to read, understand, draw out the essence and be able to deliver it to us. A primary use of this would be better search engines, but it also has other uses. Web browsers and search engines would be able to answer simple questions by mining the web for the answers. You might also be able to highlight a phrase in your browser and have it come up with more information about that phrase – such as reading an article about horses, highlighting a phrase about the proper care for a horse, and having the browser pull up more information about that subtopic.

3. Problem Definition

The main Plan is to measure based on the combination of individual similarity measures according to each of the dimensions. This combination will be produced as training across numerous observations that will affect the weight with which each dimension contributes to the final decision. Training can be performed according to different criteria. On one hand, different human subjects support their judgments on different combinations of the dimensions. On the other hand, the nature of the concept determines the most relevant dimension for each comparison[4][8]. For example, when comparing the concept scanner with the concept printer, the sort dimension could be very influential, since both are types of computer peripherals; however restrictive dimension could not be as influential because they are related to different actions. The opposite may happen with the concepts teacher and 'tutorial'. because both are related to similar actions according to the restrictive dimension, such as teaching, while the sort dimension has little influence in this case.

4. System Architecture

It's important to note that there are two different approaches to creating a semantic search engine and a semantic web. One approach is for websites to 'tag' their content, in essence telling search engines what the content is about – like a miniature book report about the web page. Another approach is for the engine to be sophisticated enough to scan through the document and figure it out on their own. Of course, this is easier said than done. The biggest fault with the idea of 'tagging' content to give search engines a hint to the content's subject is that it begs to be abused. It is a great system if everyone is using it honestly, but there are a lot of snake oil salesmen out on the web who would use any means possible to get you to their sites. But these same sites already do the same, pulling out all the stops to trick search engines into thinking their keyword phrase is the best, so we might just be par for the course on that one.

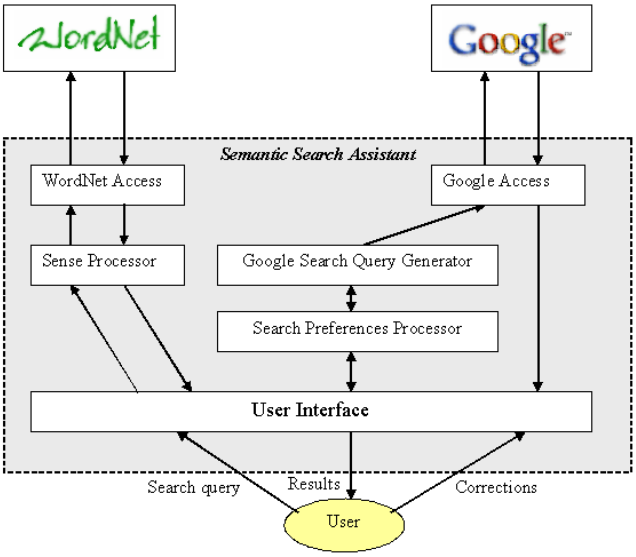


Figure 1: Semantic Search Engine Works

4.1 The Antonym Dictionary for Review Reversion

We noticed that DSA highly depends on an external antonym dictionary for review reversion.

4.1 .1 The Lexicon-based Antonym Dictionary

An example of lexicon based Antonym dictionary is WordNet. WordNet is a lexical database which groups English words into sets of synonyms called synsets, provides short, general definitions, and records the various semantic relations between these sets. Using the antonym thesaurus it is possible to obtain the words and their opposites. The WordNet antonym dictionary is simple and direct. However, in many languages other than English, such an antonym

dictionary may not be readily available. Even if we can get an antonym dictionary, it is still hard to guarantee vocabularies in the dictionary are domain consistent with our tasks. To solve this problem, we furthermore develop a corpus- based method to construct a pseudo-antonym dictionary.

5. Mathematical Model

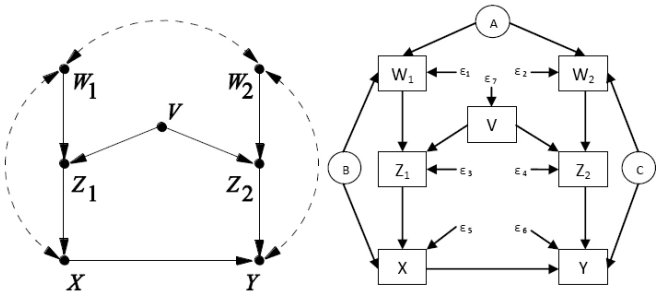


Figure 2: Mathematical Model

The main purpose of this model is a human-like interaction system is to unify the representation of each concept, relating it to the appropriate terms, as well as to other concepts with which it shares a semantic relation. Furthermore, the ontological component W1 and W2 should also be able to perform certain inferential processes, such as the calculation of semantic similarity (V) between concepts Z1 and Z2 and the output X and Y are again compared with the Component W1 and W2. The subject of similarity has been and continues to be widely studied in the fields and literature of computer science, artificial intelligence, psychology and linguistics [4]. Good similarity measures are necessary for several techniques from these fields including information retrieval, clustering, data-mining, sense disambiguation, ontology translation and automatic schema matching. The present model focuses on the study of semantic similarity between concepts in an ontology from the framework of natural interaction.

6. Existing System

1. Leacock and Chodorow (1998) also rely on the length len(c1; c2) of the shortest path between two synsets for their measure of similarity. However, they limit their attention to IS-A links and scale the path length by the overall depth D of the taxonomy:

SimLC(c1; c2) = log (len(c1; c2) / 2D)

2. Resnik: Resniks (1995) approach was, to our knowledge, the first to bring together ontology and corpus. Guided by the intuition that the similarity between a pair of concepts may be judged by the extent to which they share information, Resnik defined the similarity between two concepts lexicalized in WordNet to be the information content of their lowest super-ordinate (most specific common subsumer)

SimR(c1; c2) = log p (lso(c1; c2))

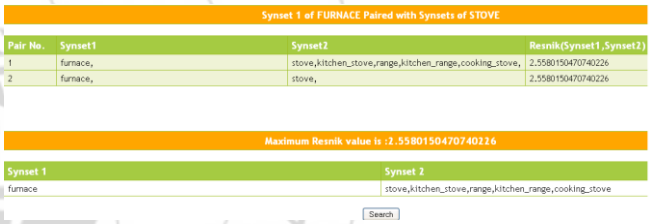
where p(c) is the probability of encountering an instance of a synset c in some specific corpus.

7. Result

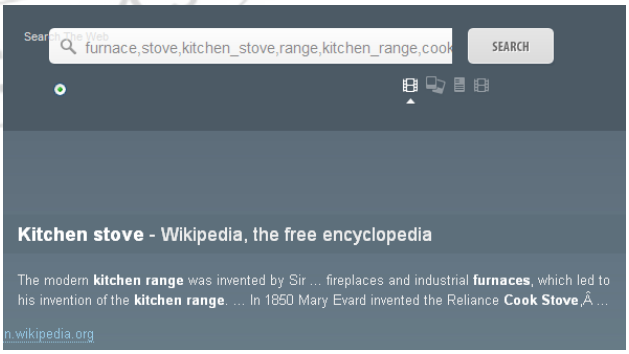
Different modules to be implemented which are part of the application are covered in the below section:

1) Login Screen: Very first screen of the system, where user needs to enter the two words for checking the similarity of two words.

2) Finding the best Synset: After clicking on the search button, the list of synset and possibility appears and which ever has the best possibility and probability will be chosen by the algorithm and leads to the search on search bar.



3) Result on Search Engine: Using the Wordnet 2.0, the database of Google is retrieved. From this the result is being provided to user. Similarly user can download the video, Images etc.



8. Conclusion

Semantic search is an upcoming technology that has set the expectations way too high. We have all been misled into thinking that these technologies are here to dethrone Google by delivering better search results. Neither of those things is true. What is true, however is that Semantic search is going to be big and it is going to help us answer questions that we simply cannot answer today - complex, inference queries

asked over the entire web as if it was a database. In contrast with these efforts, our semantic search engine, provides several means to address this issue. First, it will overcome the problem of knowledge overhead by supporting a Google-like query interface. The proposed query interface provides a simple but powerful way of specifying queries. Second, it will be able to produce precise answers or user queries by providing comprehensive means to make sense of user queries and to translate them into formal queries. In particular, the produced answers on the one hand satisfy user queries and on the other hand are self-explanatory and understandable by end users. Finally, Semantic search provides means (i.e., search renitent forms) to support end users in reformulating better queries.

References

[1] Pablo Rodriguez-Mier, Carlos Pedrinaci, Manuel Lama, and Manuel Mucientes," An Integrated Semantic Web Service Discovery and Composition Framework", IEEE Transition Paper, 10 Feb, 2015
[2] Abdeslem DENNAI, and Sidi Mohammed BENSLIMANE, "A New Measure of the Calculation of Semantic Distance between Ontology Concepts," EEDIS Laboratory, Djillali Liabes University, Sidi Bel Abbes Algeria, vol. 11(2015), pp. 48-56, July. 2015.
[3] Atul Dubey, M. Silvina Tomassone, "Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures", Department of Computer Science, University of Toronto, Toronto, Ontario, Canada Feb. 2014.
[4] Daniel Fried, and Kevin Duh, "INCORPORATING BOTH DISTRIBUTIONAL AND RELATIONAL SEMANTICS IN WORD REPRESENTATIONS", Department of Computer Science, University of Arizona, Tucson, Arizona, USA, Sep. 24, 2014.
[5] Jaap Kamps, Maarten Marx, Robert J. Mokken, Maarten de Rijke, "Using WordNet to Measure Semantic Orientations of Adjectives," Language and Inference Technology Group, ILLC, University of Amsterdam., ISSN 1424-8220, vol. 15, no. 3, pp. 311313,Mar. 2008.
[6] Julien Subercaze, Christophe Gravier, Frederique Laforest, "On metric embedding for boosting semantic similarity computations," Universite de Lyon, F-42023, Saint-Etienne, France., vol. 16, no. 3, pp. 311313, 22 Jun 2015.
[7] Vanessa Ayala-Rivera, Patrick McDonagh, Thomas Cerqueus, and Liam Murphy," Ontology-Based Quality Evaluation of Value Generalization Hierarchies for Data Anonymization," School of Computer Science and Informatics, University College Dublin, ISSN 1424-8220, vol. 17, no. 9, pp. 31846, Mar. 2008.
[8] Andrea Ballatore, Michela Bertolotto and David C. Wilson, "A Structural-Lexical Measure of Semantic Similarity for Geo-Knowledge Graphs," University of California, Santa Barbara, CA 93106, USA, ISSN 1501, vol. 19, no. 3, April. 2015.
[9] Gottfried Vossen, Miltiadis Lytras, and Nick Koudas," Editorial: Revisiting the (Machine) Semantic Web: The Missing Layers for the Human Semantic Web", IEEE Transition paper ON KNOWLEDGE AND DATA ENGINEERING, VOL. 19, NO. 2, FEBRUARY 2007,

Author Profile

Mr. Rahul Shirbhate is pursuing his Masters of Engineering in the Computer Engineering Department, Sinhgad School of Engineering, and Savitribai Phule University Pune. He received Bachelor of Technology degree in Information Technology from Govt. College of Engineering, University of Amravati, Maharashtra, India.

Prof. Vishal Mogal is the Asst. Professor of Computer Dept. at RMD SSOE College, Pune, having more than 5+ years of experience in the field of teaching and research. The domains of his research are Software Testing, Software Engineering and Web Security.