International Journal of Smart Sensor and Adhoc Network

Volume 1 | Issue 4 Article 13

April 2012

Automatic Acquisition of Similarity between Entities by Using Web Search Engine

C. Aiswarya
Department of Computer Science and Engineering, Affiliated to Anna University, Alpha college of Engineering, Thirumazhisai, Chennai, saaiswarya@gmail.com

R. Lakshmi
Department of Computer Science and Engineering, Affiliated to Anna University, Alpha college of Engineering, Thirumazhisai, Chennai., lakshmicse001@gmail.com

R. kotteswari
Department of Computer Science and Engineering, Affiliated to Anna University, Alpha college of Engineering, Thirumazhisai, Chennai., rkotteswari1805@gmail.com

Automatic Acquisition of Similarity between Entities by Using Web Search Engine

[1]C.Aiswarya ,[2]R.Lakshmi ,[3] R.kotteswari & [4] M. Antony Robert Raj

Department of Computer Science and Engineering, Affiliated to Anna University, Alpha college of Engineering, Thirumazhisai, Chennai.
Email: [1]saaiswarya@gmail.com,[2] lakshmicse001@gmail.com & [3]rkotteswari1805@gmail.com

Abstract

Web mining is the application of data mining technology to discover patterns from the web. The various tasks on web such as relation extraction, community mining, document clustering and automatic metadata extraction. A previously proposed web-based semantic similarity measures on three benchmark datasets showing high correlation with human rating. One of the main problems in information retrieval is to retrieve a set of documents that is semantically related to given user query. We propose an automatic acquisition method to estimate semantic relation between two words by using pattern extraction algorithm and sequential clustering algorithm.

Keywords - web mining, information extraction, text snippets, and page counts.

1. Introduction

Automatic acquisition of similarity between entities by using web search engine is an important problem in web mining. A web search engine is designed to search for information on the World Wide Web and FTP services. The search results are generally presented in a list of results often referred to as SERPS, or "search engine results pages". Some search engines also mine data available in databases or open directories. Unlike web directories which are maintained by human editors, search engine maintains real-time information by running an algorithm on a web crawler.

Page counts and snippets are two useful information sources provided by most web search engines. Page count of a query is an estimate of the number of pages that contain the query words. In general page count may not necessarily by equal to the word frequency because the query word might appear many times on one page. Snippets, a brief window of text extracted by a search engine around the query term in a document, provide useful information regarding the local context of the query term. Semantic similarity measures defined over snippets have been used in query expansion, personal name, disambiguation and community mining. Processing of snippets is also efficient because it obviates the trouble of downloading web pages, which might be time consuming depending on the size of the pages. A widely acknowledged drawback of using snippets is that, because of huge scale of the web and the large number of documents in the result set, only those snippets for the top ranking results for a query can be processed efficiently.

1.1. Motivation

The motivation of this paper is present an automatically extracted lexical pattern based approach to compute the semantic similarity between words or entities using text snippets retrieved from a web search engine. We propose a lexical pattern extraction algorithm that considers word subsequences in text snippets. Moreover, the extracted sets of pattern are clustered to identify the different patters that describe the same semantic relation.

1.2. **Web Search Engine**

How web search engine work

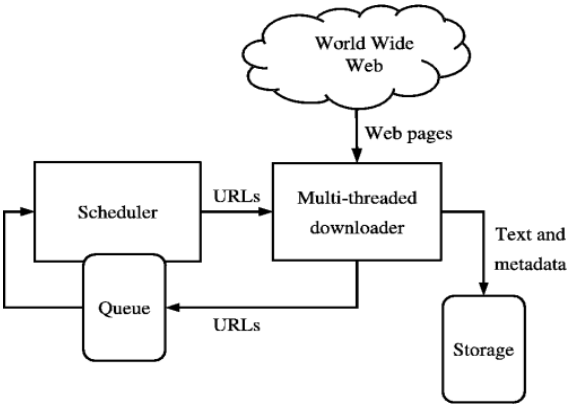

Fig 1.1 OVERVIEW OF WEB SEARCH ENGINE

Web Search engine work by storing information about many web pages, which they retrieve from the html itself. These pages are retrieved by a web crawler. (Sometime also
Known as spider)-an automated web browser which follows every link on the site. Exclusions can be made by the use of robots.txt. The content of each page are then analyzed to determine how it should be indexed (for example, words are extracted from the titles, headings or special fields called Meta tags). When user enters a query into a search engine (typically by using keywords) the engine examines its index and provides a listing of best –matching WebPages accordingly to its criteria, usually with a short summary containing the documents title and sometime parts of the text. The index is built
From the information is indexed unfortunately, there are currently no known public search
Engines that allows documents to be searched by data. most search engines support the use of the Boolean operators AND, OR and NOT to further specify the search query. We propose method a method that considers both page counts and lexico-syntactic pattern extracted from snippets, there by overcoming the problems described above. for example, let us consider the following snippets from Google for the query "Apple" AND "compute" Boolean operators are for literal searches that allows the user to refine and extend the term of the search. The engine looks for the words or phrases exactly as entered. Some search engine provides an advanced feature called proximity search which allows users to define the distance between keywords.

2 **Related Works**

Given taxonomy of words, a straight forward method to calculate similarity two words to find length of shortest path connecting two words in taxonomy [7]. If a word polysemous, then the multiple path exits between two words. In such cases, only shortest path between any two senses of the word is considered for calculating similarity. A problem that is frequently acknowledged with this approach is that it relies on the notion that all links in taxonomy represent a uniform distance. However, semantic relation can be expressed using more than one lexical pattern that conveys same semantic relation that enables us to represent semantic similarity between two words accurately. For this purpose sequential pattern clustering. Both page count similarity scores and lexical pattern are used to define various features to represent relation between two words. Using this feature representation of word-pair, we train two-class support vector machine.

3 Method

3.1 Outline

The search engine looks in the index file for matches and gathering matching pages, sorted by relevance formats the results listing page send the result page back to the user.

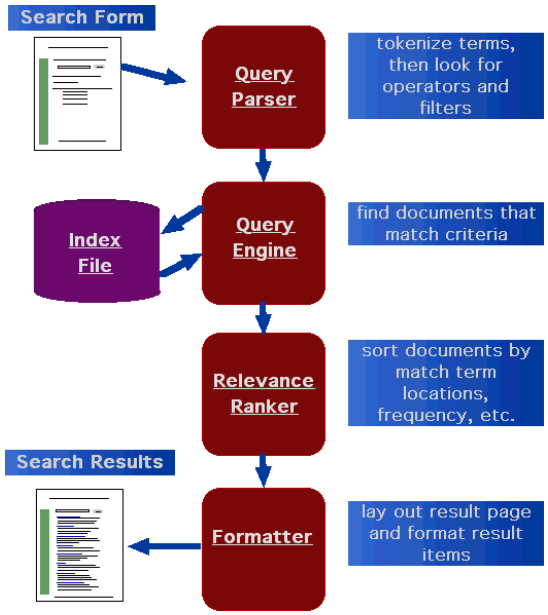

Fig 1.2 Overall Architecture

3.2 **Lexical Pattern Extraction**

The World Wide Web and its associated distributed information services, such as yahoo, google, America on-line which provide world –wide, on-line information services, data objects are linked to facilities the interactive access. (For example- a web search based on single keyword may return hundreds of web page pointers containing the keyword, but most of the pointers will be very weekly related to what the user wants to find)
Web search engines provide an efficient interface to this vast information such as page counts and snippets are to useful information sources provided by most web search engines. Page count of a query is an estimate of the number of pages that contain the query words. In general, page count may not necessarily be equal to the word frequency because the query word might appear many times on one page so, we present an automatically extracted lexical syntactic patterns-based approach to compute the semantic similarity

between words or entities using text snippets retrieved from a web search engine.

3.3 **Lexical Pattern Clustering**

A cluster is therefore a collection of objects which are "similar" between them and are "dissimilar" to the objects belonging to the other clusters. A semantic relation can be expressed using more than one pattern. For example, consider the two distinct patterns X is a Y, and X is a large Y. But these patterns indicate that there exits and is-a relation between X and Y. Identifying the different patterns that express the same semantic relation enables us to represent the relation between two words accurately. According to the distributional hypothesis has been used in various related tasks such as identifying related words, and extracting paraphrases. If we consider the word pairs that satisfy (i.e., co-occur with) a particular lexical patterns as a context of a lexical pair, then from the distributional hypothesis, it follows that the lexical patterns which are similarly distributed over word pairs must be semantically similar.

3.4 **Measuring Semantic Similarity**

We describe the machine learning approach to combine both page counts based co-occurrence measures, and snippets-based lexical pattern to construct a robust semantic similarity measure.

3.5 **Ranking Search Results**

An automatic method is described to estimate the semantic similarity between words or entities using web search engines with ranking the search results occur. Accurately measuring the semantic similarity between words is an important problem in web mining, information retrieval, and natural language processing. Web mining applications such as, community extraction, relation detection, and entity disambiguation; require ability to accurately measure the semantic similarity between concepts or entities. Based on the user given search keyword the ranking takes place.

4. **Conclusion**

We proposed a semantity similarity measure using both page counts and snippets retrieved from a web search engine for two given words or named entities. The method consists of four page-count-based similarity scores and automatically extracted lexico-syntactic patterns. We integrated page counts based similarity scores with lexico-syntactic patterns using support vector machines. Training data were automatically generated using wordnet synsets. Proposed method outperformed all the baselines including previously proposed web based semantic similarity measures on a benchmark datasets. Which is efficient and scalable because it only processes the snippets (no downloading of web pages is necessary) for the top ranking results by google. Results of our experiments indicate that the proposed method can robustly capture semantic similarity between named entities. In future research, we intend to apply the proposed semantic similarity measures in automatic synonym extraction, query suggestion and name alias recognition.

Reference

[1] A. Kilgarriff, "Googleology Is Bad Science," Computational Linguistics, vol. 33, pp. 147-151, 2007.

[2] M. Sahami and T. Heilman, "A Web-Based Kernel Function for Measuring the Similarity of Short Text Snippets," Proc. 15th Int'l World Wide Web Conf., 2006.

[3] D. Bollegala, Y. Matsuo, and M. Ishizuka, "Disambiguating Personal Names on the Web Using Automatically Extracted Key Phrases," Proc. 17th European Conf. Artificial Intelligence, pp. 553-557, 2006.

[4] H. Chen, M. Lin, and Y. Wei, "Novel Association Measures Using Web Search with Double Checking," Proc. 21st Int'l Conf. Computational Linguistics and 44th Ann. Meeting of the Assoc. for Computational Linguistics (COLING/ACL '06), pp. 1009-1016, 2006.

[5] M. Hearst, "Automatic Acquisition of Hyponyms from Large Text Corpora," Proc. 14th Conf. Computational Linguistics (COLING), pp. 539-545, 1992..

[6] M. Pasca, D. Lin, J. Bigham, A. Lifchits, and A. Jain, "Organizing and Searching the World Wide Web of Facts - Step One: The One-Million Fact Extraction Challenge," Proc. Nat'l Conf. Artificial Intelligence (AAAI '06), 2006.

[7] R. Rada, H. Mili, E. Bichnell, and M. Blettner, "Development and Application of a Metric on Semantic Nets," IEEE Trans. Systems, Man and Cybernetics, vol. 19, no. 1, pp. 17-30, Jan./Feb. 1989.

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

[9] D. Mclean, Y. Li, and Z.A. Bandar, "An Approach for Measuring Semantic Similarity between Words Using Multiple Information Sources," IEEE Trans. Knowledge and Data Eng., vol. 15, no. 4, pp. 871-882, July/Aug. 2003.

[10] G. Miller and W. Charles, "Contextual Correlates of Semantic Similarity," Language and Cognitive Processes, vol. 6, no. 1, pp. 1-28, 1998

[11] D. Lin, "An Information-Theoretic Definition of Similarity," Proc. 15th Int'l Conf. Machine Learning (ICML), pp. 296-304, 1998

[12] R. Cilibrasi and P. Vitanyi, "The Google Similarity Distance," IEEE Trans. Knowledge and Data Eng., vol. 19, no. 3, pp. 370-383, Mar. 2007.

[13] M. Li, X. Chen, X. Li, B. Ma, and P. Vitanyi, "The Similarity Metric," IEEE Trans. Information Theory, vol. 50, no. 12, pp. 3250-3264, Dec. 2004.

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

[15] R. Rosenfield, "A Maximum Entropy Approach to Adaptive Statistical Modelling," Computer Speech and Language, vol. 10, pp. 187-228, 1996.

[16] D. Lin, "Automatic Retrieval and Clustering of Similar Words," Proc. 17th Int'l Conf. Computational Linguistics (COLING), pp. 768-774, 1998.

[17] J. Curran, "Ensemble Methods for Automatic Thesaurus Extraction," Proc. ACL-02 Conf. Empirical Methods in Natural Language Processing (EMNLP), 2002.

[18] C. Buckley, G. Salton, J. Allan, and A. Singhal, "Automatic Query Expansion Using Smart: Trec 3," Proc. Third Text REtreival Conf., pp. 69-80, 1994.

[19] V. Vapnik, Statistical Learning Theory. Wiley, 1998. [20] K. Church and P. Hanks, "Word Association Norms, Mutual Information and Lexicography," Computational Linguistics, vol. 16, pp. 22-29, 1991.

[21] Z. Bar-Yossef and M. Gurevich, "Random Sampling from a Search Engine's Index," Proc. 15th Int'l World Wide Web Conf., 2006.

[22] F. Keller and M. Lapata, "Using the Web to Obtain Frequencies for Unseen Bigrams," Computational Linguistics, vol. 29, no. 3, pp. 459- 484, 2003.

[23] M. Lapata and F. Keller, "Web-Based Models for Natural Language Processing," ACM Trans. Speech and Language Processing, vol. 2, no. 1, pp. 1-31, 2005.