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


## ABSTRACT

Measuring the semantic similarity between words is an important component in various tasks on the web such as relation extraction, community mining, document clustering, and automatic metadata extraction. Despite the usefulness of semantic similarity measures in these applications, accurately measuring semantic similarity between two words (or entities) remains a challenging task. We propose an empirical method to estimate semantic similarity using page counts and text snippets retrieved from a web search engine for two words. Specifically, we define various word co-occurrence measures using page counts and integrate those with lexical patterns extracted from text snippets. To identify the numerous semantic relations that exist between two given words, we propose a novel pattern extraction algorithm and a pattern clustering algorithm. The optimal combination of page counts-based co-occurrence measures and lexical pattern clusters is learned using support vector machines. The proposed method outperforms various baselines and previously proposed web-based semantic similarity measures on three benchmark data sets showing a high correlation with human ratings. Moreover, the proposed method significantly improves the accuracy in a community mining task.


## EXISTING WORK:

- 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 the ability to accurately measure the semantic similarity between concepts or entities.

- In information retrieval, one of the main problems is to retrieve a set of documents that is semantically related to a given user query.

- Efficient estimation of semantic similarity between words is critical for various natural language processing tasks such as word sense disambiguation (WSD), textual entailment, and automatic text summarization.

- 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.

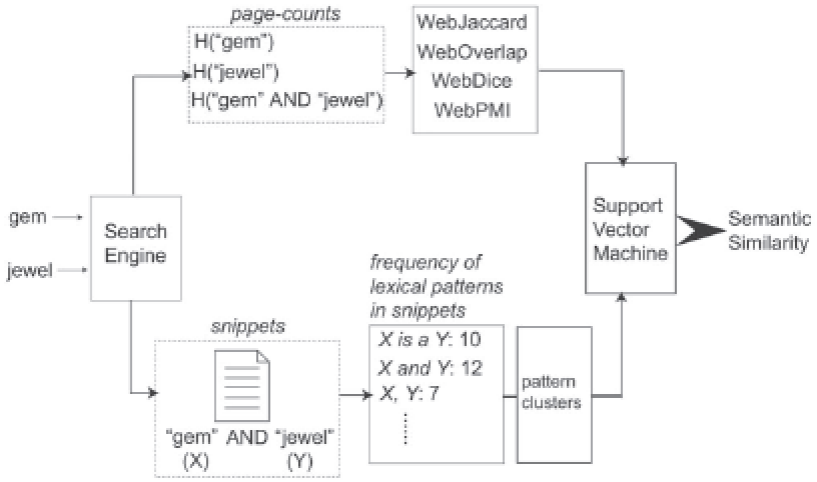
## PROPOSED WORK:

- We propose an automatic method to estimate the semantic similarity between words or entities using web search engines.

- Web search engines provide an efficient interface to this vast information. 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 be equal to the word frequency because the queried word might appear many times on one page.

- 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.




## HARDWARE REQUIREMENTS

| | | |
|---|---|---|
| PROCESSOR | : | PENTIUM 4 CPU 2.40GHZ |
| RAM | : | 128 MB |
| HARD DISK | : | 40 GB |
| KEYBOARD | : | STANDARD |
| MONITOR | : | 15" |

## SOFTWARE REQUIREMENTS

| | | |
|---|---|---|
| FRONT END | : | C#.NET |
| OPERATING SYSTEM | : | WINDOWS XP |
| DOCUMENTATION | : | MS-OFFICE 2007 |

## MODULES:

- Lexical Pattern Extraction
- Lexical Pattern Clustering
- Measuring Semantic Similarity
- Ranking search results

## MODULES DESCRIPTION:

### Lexical Pattern Extraction

In this module, Words in Page are extracted. It uses counts-based co-occurrence measures. Lexical Pattern Clustering. This can be problematic if one or both words are polysemous, or when page counts are unreliable. On the other hand, the snippets returned by a search engine for the conjunctive query of two words provide useful clues related to the semantic relations that exist between two words. A snippet contains a window of text selected from a document that includes the queried words. Snippets are useful for search because, most of the time, a user can read the snippet and decide whether a particular search result is relevant, without even opening the url. Using snippets as contexts is also computationally efficient because it obviates the need to download the source documents from the web, which can be time consuming if a document is large.

### Lexical Pattern Clustering

Typically, 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. Both these patterns indicate that there exists 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, words that occur in the same context have similar meanings. 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 pattern as the context of that lexical pair, then from the distributional hypothesis, it follows that the lexical patterns which are similarly distributed over word pairs must be semantically similar.

**Measuring Semantic Similarity**

We defined four co-occurrence measures using page counts. We showed how to extract clusters of lexical patterns from snippets to represent numerous semantic relations that exist between two words. In this module, we describe a machine learning approach to combine both page counts-based co-occurrence measures, and snippets-based lexical pattern clusters to construct a robust semantic similarity measure.

**Ranking search results**

In this module, an automatic method 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 the ability to accurately measure the semantic similarity between concepts or entities. Based on the similarity between the user given search keyword, the ranking takes place.

## REFERENCE:

Danushka Bollegala, Yutaka Matsuo, and Mitsuru Ishizuka, "A Web Search Eninge-Based Approach to Measure Semantic Similarity between Words", **IEEE Transactions on Knowledge and Data Engineering, Vol.23, 2011.**