Measuring semantic similarity between words using Web search engines

VARIKUTI NARESH,
Associate Professor,
Department of Computer Science and Engineering,

SRINIVAS GADARI,
Associate Professor,
Department of Electronics and Communications Engineering,

Siddhartha Institute of Technology and Sciences,
Narapally, Hyderabad, Telangana – 500 088.

Abstract

An internet search engine developed an empirical approach for detecting semantic similarity based on page counts and text samples from two queries. It achieves this by combining lexical patterns taken from text excerpts with page counts to provide many word co-occurrence measures. A unique pattern extraction approach and a pattern clustering algorithm are developed to find the numerous semantic links that exist between two provided words. To discover the best mix of co-occurrence metrics based on page counts and lexical pattern clusters, Sequential Minimal Optimization is utilised. In a group mining operation, the proposed approach considerably enhances accuracy. The proposed semantic similarity metric is utilised in a community extraction job to uncover linkages between items, particularly people. With statistically significant accuracy and recall values, the suggested technique outperforms the baselines. The results of the community mining challenge indicate that the suggested method may be used to compare the semantic similarity of not just words, but also named items for which human lexical ontologies are either missing or incomplete.

1. Introduction

Web search engines give a quick and easy way to access all of this data. Most web search engines give two useful information sources: page counts and snippets. A query's page count is an estimate of how many pages include the query terms. Because the same word may appear many times on a single page, page count may not always be equal to word frequency. As a result, the research provides an approach that takes into account both page counts and lexical syntactic patterns retrieved from snippets, and shows that it can solve the challenges outlined above.

In web mining, information retrieval, and natural language processing, accurately quantifying semantic similarity between words is a major challenge. The capacity to reliably quantify the semantic similarity of ideas or entities is required for web mining applications such as community extraction, connection identification, and entity disambiguation. One of the most difficult challenges in information retrieval is retrieving a group of documents that are semantically relevant to a given user query.

The criterion for determining similarity is implementation dependant. Data Clustering is a strategy for physically storing information that is conceptually comparable. The amount of disc accesses should be

reduced in order to improve database system efficiency. Objects with identical attributes are grouped together in one class, and a single disc access makes the entire class available.

The real data mining work is the automated or semi-automated processing of vast amounts of data in order to identify previously unknown interesting patterns such as clusters of data records (cluster analysis), anomalies (anomaly identification), and relationships (association rule mining). This often entails the use of database techniques like spatial indexes.

2. Literature survey

When you search the web for data about a certain individual, a search engine returns multiple pages. Some of these sites may be for persons with the same name. How will they distinguish between these different people with the same name? This study provides an unsupervised method that generates unique words to disambiguate distinct persons with the same name (i.e. namesakes).

They may use the huge amount of publications on the web to identify additional context for a brief text excerpt to assist us in achieving this aim. We may match the user's first inquiry against a big collection of prior user requests using their short text similarity kernel to discover further similar queries to offer to the user. As a consequence, the results of the similarity function may be used directly in an end-user application.

However, employing typical document similarity techniques, such as the commonly used cosine coefficient, to such brief text fragments frequently yields insufficient results. Indeed, in each of the above cases, using the cosine would result in a similarity of 0 because each text pair has no common phrases. Even though two snippets share terminology, they may use the phrase in distinct contexts.

In text analysis, there are a variety of situations when we wish to see how similar two short snippets are. They want to determine if there is a high amount of semantic similarity between two text samples, such as "United Nations Secretary-General" and "KofiAnnan." Similarly, the terms "AI" and "Artificial Intelligence" have a lot in common, even if they don't have the same exact meaning.

Web search engines offer a convenient way to access this large amount of information. Most web search engines include page counts and snippets as helpful information sources. A query's page count is an estimate of how many pages include the query terms. In general, the page count may not be equal to the word frequency because the searched term may appear many times on a single page. As a result, the research recommends an approach that takes into account both page counts and lexical syntactic patterns retrieved from samples.

3. Methodology

The process of turning user-generated inputs into a computer-readable format is known as input design. Input design is one of the most expensive parts of a computerised system's functioning, and it's frequently the system's biggest issue. The design and manner of fault input may generally be traced back to a wide variety of system issues. Every step of the input design process should be meticulously studied and planned.

The system receives user input, processes it, and generates a result. Input design is the connection that connects the information system to its users' reality. To provide suitable information to the user, the system must be user-friendly. During the input design phase, decisions are taken.

- To provide a cost-effective input technique.
- To attain the best level of precision feasible.
- To verify that the user comprehends the input.

A system's input data does not always have to be raw data that was entered into the system from the beginning. Another system's or subsystem's output can likewise be used. The design of input encompasses all stages of data entry, from the development of original data through the actual entry of data into the system for processing. Identifying the data required, establishing the properties of each data item, capturing and preparing data for computer processing, and assuring data accuracy are all part of the input design process.

After the system's management was accepted, the system was installed in the company, originally running in tandem with the company's existing manual system. The system was tested using real-time data and proven to be error-free and user-friendly. When the system performed the original design, implementation is the process of transforming a new or changed system design into an operational one; a demonstration of the working system was provided to the end user. By introducing multiple permutations of test data into the system, this procedure is used to check and identify any logical snafu in the system's operation.

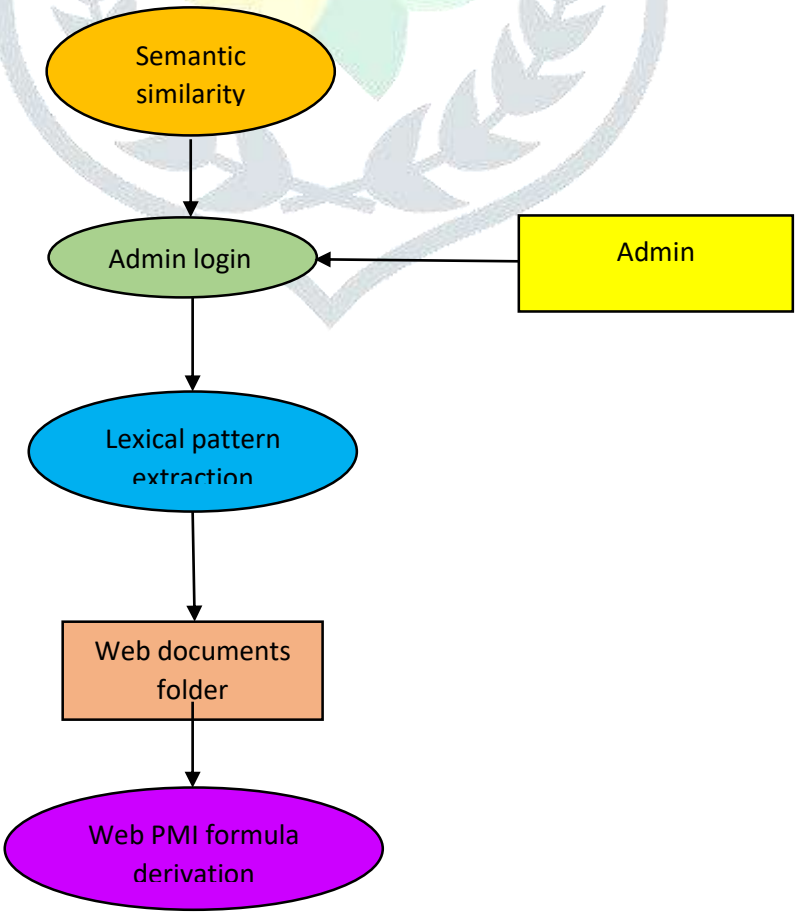

Fig 1. Data Flow Diagram

Since the results seem to be the most essential informants for users, a better design should enhance the system's interactions with them and aid decision-making. The way output is presented and the arrangement available for gathering knowledge are both elaborated by form design.

4. Result and discussion

Each module is thoroughly tested. After all of the modules have been tested, the modules are combined, and the completed system is tested using test data that has been particularly prepared to demonstrate that the system will work effectively in all of its features. As a result, system testing serves as a validation that everything is in order as well as a chance to demonstrate to the user that the system works. Only acknowledged orders are taken for consolidation after the modules have been checked. The module has numerous fields in which the input string must be legitimate, the maximum character must be verified, and a notification stating "should not exceed the limit" will be displayed. Validation testing code is used in this project to determine whether the provided input is legitimate or not.

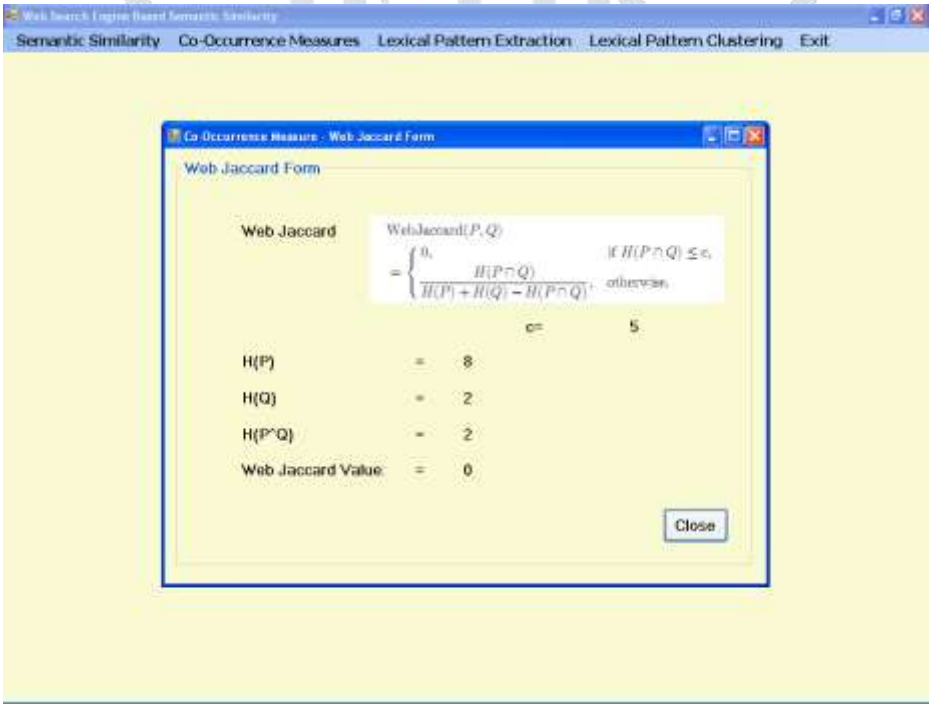

Fig. 2 Web Jaccard

The co-occurrence value, determined using the online Jaccard technique, is displayed in this manner.

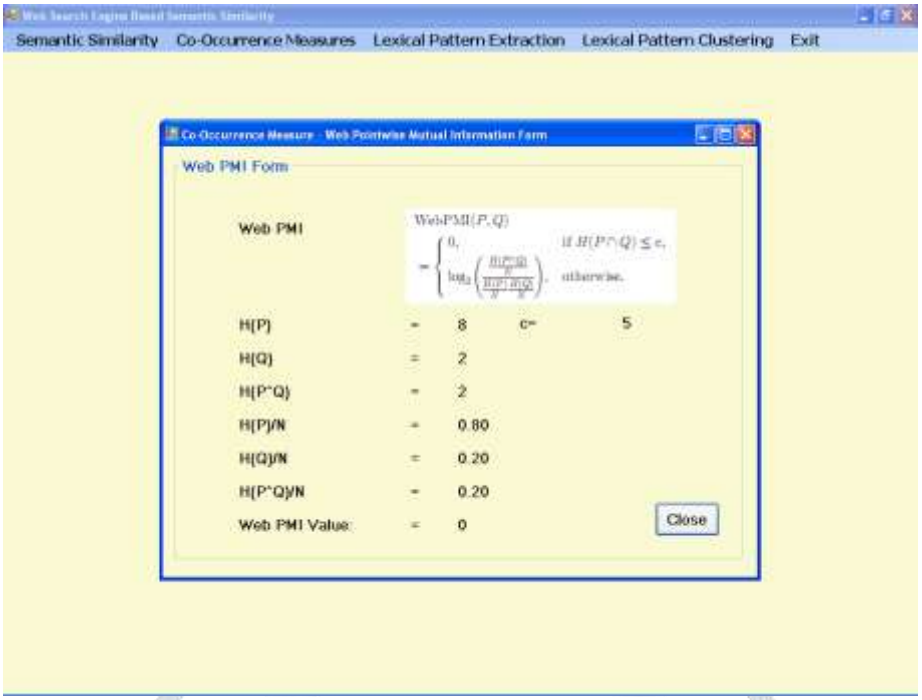

Fig 3. Web PMI

The co-occurrence value, determined using the web PMI approach, is displayed in this manner.

5. Conclusion

A sequential pattern clustering approach was created in order to locate distinct lexical patterns that indicate the same semantic link. It creates different word co-occurrence metrics by combining lexical patterns taken from text excerpts with page counts. The attributes of a word pair were determined using both page counts-based co-occurrence metrics and lexical pattern clusters. Sequential Minimal Optimization was trained using attributes obtained from WordNet synsets for synonymous and non-synonymous word pairings. For two terms, the researchers presented a semantic similarity score based on page counts and excerpts from an internet search engine. Using page counts, four word co-occurrence metrics were calculated. It developed lexical pattern extraction, a method for extracting a large number of semantic links between two words.

References

1. Bollegala.D, Matsuo.Y, and Ishizuka.M,(2006) “Disambiguating Personal Names on the Web Using Automatically Extracted Key Phrases,” Proc. 17th European Conf. Artificial Intelligence, pp. 553-557.
2. Church.K and Hanks.P,(1991) “Word Association Norms, Mutual Information and Lexicography,” Computational Linguistics, vol. 16, pp. 22-29.
3. Cilibrasi.R and Vitanyi.P,(2007) “The Google Similarity Distance,” IEEE Trans. Knowledge and Data Eng., vol. 19, no. 3, pp. 370-383.
4. Cristianini.N and Shawe-Taylor.J.(2000) An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. Cambridge University Press.

5. Gabrilovich.E and Markovitch.S,(2007), "Computing Semantic Relatedness Using Wikipedia-Based Explicit Semantic Analysis," Proc.Int'l Joint Conf. Artificial Intelligence (IJCAI '07), pp. 1606- 1611.
6. Pasca.M, Lin.D, Bigham.J, Lifchits.A, and Jain.A,(2006) "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).
7. Sahami.M and Heilman.T,(2006) "A Web-Based Kernel Function for Measuring the Similarity of Short Text Snippets," Proc. 15th Int'lWorld Wide Web Conf.
8. Strube.M and Ponzetto.S.P,(2006) "Wikirelate! Computing Semantic Relatedness Using Wikipedia," Proc. Nat'l Conf. Artificial Intelligence (AAAI '06), pp. 1419-1424.
9. Ted Pedersen, Amruta Purandare, and Anagha Kulkarni,(2005) 'Name discrimination by clustering similar contexts', in Proceedings of the Sixth International Conference on Intelligent Text Processing and Computational Linguistics.
10. Xin Li, Paul Morie, and Dan Roth,(2005) 'Semantic integration in text, from ambiguous names to identifiable entities', AI Magazine, American Association for Artificial Intelligence, Spring, 45–58.