International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 2 Issue 2, February- 2013

A Semantic Similarity Measure Using Both Page Counts And Snippets Retrieval From A Web Search Engine

Prof(Dr).V. Saravanan
Professor & Director
Department of Computer Applications
Sri Venkateswara College of Computer Applications and Management

Mrs.R.Kousalya,Ph.d.,Scholar,
Manonmaniam Sundaranar University,
Head of the Department,
Department of Computer Applications,
Dr. N.G.P Arts and Science College, Coimbatore.

K.Kalaivani,
M.Phil. Scholar
Dr. N.G.P Arts and Science College,Coimbatore.

Abstract

This web search engine provides the most semantic relativity between the given words, and it will generate the semantic measures automatically. This kind of extraction improves the efficiency of the user search. It is an automatic method to estimate the semantic similarity between words or entities using web search engines text snippets and a lexical pattern extraction algorithm that considers word subsequences in text snippets. It is time consuming to analyze each document separately. Web search engines provide an efficient interface to vast information. Page counts and snippets are two useful information sources provided by most web search engines. And then train a two-class support vector machine to classify synonymous and non synonymous word pairs. Both novel pattern extraction algorithm and pattern clustering algorithm outperforms well in the case of page counts for given words with the text snippets.

Index terms

Web Content Mining,Semantic Web,Web Structure Mining,Page Ranking,Pattern Clustering

Introduction

Search engines have become the most helpful tool for obtaining useful information from the Internet. However, the search results returned by even the most popular search engines are not satisfactory. It is not uncommon that search engines return a lot of Web page links that have nothing to do measuring the semantic similarity between words is an important component in various tasks on the web such as relation extraction, content mining, document clustering, and automatic metadata extraction. In Semantic Web, the semantics information is presented by the relation with others and is recorded from the effective content and data retrieval. The retrieval process is more important and that should be effectively done based on the similarity. The similarity measures should concentrate on both data extraction and filtering of those data for effective ranking.

As we have experience in using well-known search engines every day, the result set returned by search engines is really too big and is mostly useless. We have to continually click the “next page” to obtain the Web pages users really want. The reason is that, when the user wants to search some information in the Web, the search engine abstracts the information to the keyword combination and then submits it. The relationship between keywords is obvious to users, while it is not for search engines. If the Web page only includes the keywords and there is no relationship between keywords in the context of the Web page, the Web page does not provide what the user wants. In this case, we say the Web page is a keywords-isolated page. However, there are many keywords-isolated pages in the result set returned by traditional search engines. In fact, because of the constraints of the current Web architecture, search engines cannot exclude these keywords-isolated.

Problem Statement

A common technique that is used to find template is alignment: either string alignment or tree alignment. As for the problem of distinguishing template and data, most approaches assume that HTML tags are part of the template, while the existing systems considers a general model where word tokens can also be part of the template and tag tokens can also be data. When the template differs the extraction and further process will not be efficient.

The main objective is to overcome the above problem and providing effective ranking techniques based on the user frequent access patterns.

Related work

The existing system uses a straightforward method to calculate similarity between two words is to find the length of the shortest path connecting the two words in the taxonomy. A problem that is frequently acknowledged with this approach is that it relies on the notion that all links in the taxonomy represent a uniform distance. These processes are just like a manual system process, not extracting automatically. If the measures are based on the shortest path then how can retrieve the most related items? It will produce only the relevant results matches to the user query. It uses the page counts to retrieve results but using page counts alone as a measure of co-occurrence of two words presents several drawbacks. The page count analysis ignores the position of a word in a page; page count of a polysemous word might contain a combination of all its senses. This system is time consuming depending on the size of the pages. Therefore, no guarantee exists that all the information we need to measure semantic similarity between a given pair of words is contained in the top-ranking snippets.

FRAMEWORK OF WEB STRUCTURE MINING

Information Retrieval on the Web

Definition: Information retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text, but also images, videos, music ...) that satisfies an information need from within large collections (usually stored on computers).

For decades information retrieval was used by professional searchers, but nowadays hundreds of millions of people use information retrieval daily. The field of IR also covers document clustering and document classification. Given a set of documents, clustering is the task of coming up with a good grouping of the documents based on their contents. Given a set of topics, and a set of documents, classification is the task of assigning each document to its most suitable topics, if any. IR systems can also be classified by the scale on which they operate. Three main scales exist:

- IR on the web.
- IR on the documents of an enterprise.
- IR on a personal computer.

When doing IR on the web, the IR system will have to retrieve information from billions of documents. Furthermore, the IR system will have to be aware of some webs, where its owners will manipulate it, so that their web can appear on the top results for some specific searches. Moreover, the indexing will have to filter, and index only the most important information, as it is impossible to store everything.

Boolean Retrieval Model

Boolean queries only can express the appearance or not appearance of some terms in a document. This model of queries is very limited, and cannot rank the results: a document satisfies the query or it does not, but there is no middle course.

Extended Boolean Model is similar to the Boolean Retrieval Model, but with some additional operators as term proximity operators. The Extended Boolean Model was the most used during the early 90's.

Ranked Retrieval Model

Ranked Retrieval Model is more complex than the Boolean Retrieval Model, and allows the user to execute queries in free text (without boolean or proximity operators). This feature makes Ranked Retrieval Model more user-friendly than Boolean Retrieval Model and Extended Boolean Model. Furthermore, the results of the search are ranked by score, so that the most representative documents of the search will appear on the top of the results.

Therefore search engines also allow the execution of boolean queries when using the “Advanced Search" option, as using boolean operators in the queries can help to get a more selective result. This makes boolean queries specially useful when the user knows what he/she is looking for.

Crawling

A search engine needs to have an index containing information about a set of web pages. Before indexing the documents, I need to have the documents. The component that will provide the documents and their content is the crawler.

The crawler will surf the Internet, or a part of it, searching for the most interesting web pages. The interesting pages will be stored locally, so that they can be indexed later. The crawler is also known as bot or spider.

In this case, I have a focused crawler that uses information from the user to focus the crawling.

Page Ranking

The purpose of Page Ranking is to measure the relative importance of the pages in the web. There are many algorithms for this purpose.

The most important ones are: Hyper Search, Hyperlink-Induced Topic Search (HITS), PageRank, Trust Rank, and OPIC.

Hyper Search

Hyper Search has been the first published technique to measure the importance of the pages in the web. This algorithm served as a base for the next ones.

Hyperlink-Induced Topic Search

HITS algorithm, also known as Hubs and Authorities, is a link analysis algorithm for the web. It is executed at query time and is used to modify the ranking of the results of a search by analyzing the link structure of the pages that will appear in the result of the search.

HITS algorithm assigns two different values to each web page: its authority value, and its hub value. The authority value of a page represents the value of the content in the page, meanwhile the hub value estimates the value of its links to other pages.

The first step in the HITS algorithm is to retrieve the set of pages in the result of the search, as the HITS algorithm only analyzes the structure of the pages in the output of the search, instead of all the web pages.

Page Rank

Page Rank is a link analysis algorithm to measure the page relevance in a hyperlinked set of documents, such as the World Wide Web. This algorithm assigns a numerical weight to each document.

This numerical weight is also called PageRank of the document. The PageRank of a web page represents the likelihood that a person randomly clicking will arrive at this page. The PageRank algorithm requires several iterations to be executed.

At each iteration, the values will be better approximated to the real value. In its simplest form, PageRank uses the next formula for each web page at each iteration:

$$PR(u) = \sum_{v \in B_u} \frac{PR(v)}{L(v)}$$

Where u is a web page, Bu is the set of pages that link to u, PR (u) is the PageRank of u, and L (u) is the number of out-links in page u.

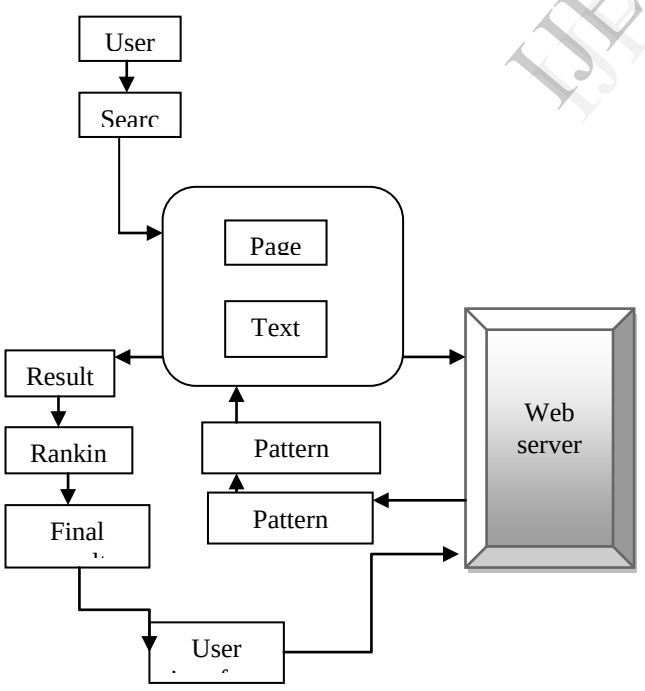
At each iteration, the PR (u) of each page u will be updated according to the values of PR (u) in the latest iteration. After several iterations, the value contained inPR (u) will be a good approximation to its real value.

For some weird structures of the links, the PageRank algorithm explained above may not converge or may have several different solutions. Solutions to this problem exist, but they are not explained in this thesis, as they are out of the scope of the thesis.

IMPLEMENTATION STEPS & METHODOLOGY

PATTERN EXTRACTION

The input texts are then split for page count calculation, this will helps to get the related items of those inputs as individual and also in combine. After that the text snippets, which are used to retrieve the related item from the online will get and make an extraction from that too.



Architecture design

Snippets

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.

Preprocessing

Preprocessing is an important step for mining tasks, whereby, the features of a data set are modified so as to make information extraction reliable and convenient. Preprocessing is necessitated due to one or more of the following reasons.

• Presence of noise in data: noise may disturb the information extraction process by making the data less than ideal.

• sparsity of data: this results in a lack of information regarding certain portions of the data space, and consequently, inference cannot be generalized easily to unseen examples

- Text based Preprocessing
- Link based Preprocessing
- Query terms
- URL
- Rank

preprocessing these items for getting clean data.

Semantic Web Search

The aim of this paper is to show how to make use of relations in Semantic Web page annotations with the aim of generating an ordered result set, where pages that best fit, the user query is

displayed first. The ideas of exploiting ontology-based annotations for information retrieval are considered to play a key role in the Semantic Web. In fact, it has been recently outlined that in order to fully benefit on semantic contents, a way for achieving relation based ranking has to be found.

A traditional search engine like Google would return both pages without considering the information provided by the semantic mark. On the other hand, a semantic search engine would take into account keyword concept Associations and would return a page only if keywords (or synonyms, homonyms, etc.) are present within the page and they are related to associated concepts. Finally, a relation-based search engine like the one presented would go beyond pure "keyword isolated" search and would include these pages in the result set only if there exist enough relations linking considered concepts.

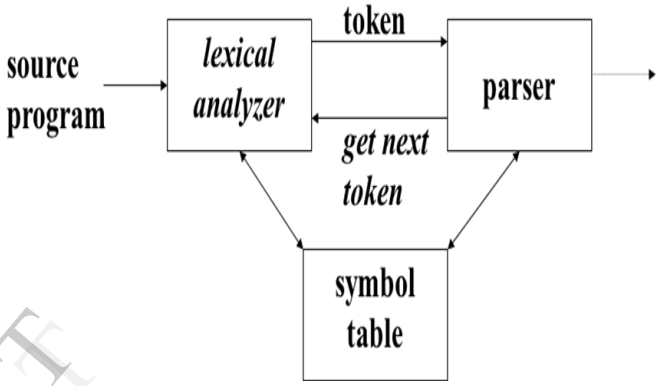
The Web has become the world's largest database, with search being the main tool that allows organizations and individuals to exploit its huge amount of information. Search on the Web has been traditionally based on textual and structural similarities, ignoring to a large degree the semantic dimension, i.e., understanding the meaning of the query and of the document content. Combining search and semantics gives birth to the idea of semantic search. Traditional search engines have already advertised some semantic dimensions. Some of them, for instance, can enhance their generated result sets with documents that are semantically related to the query terms even though they may not include these terms. Nevertheless, the exploitation of the semantic search has not yet reached its full potential.

Semantic search seeks to improve search accuracy by understanding searcher intent and the contextual meaning of terms as they appear in the searchable dataspace, whether on the Web or within a closed system, to generate more relevant results.

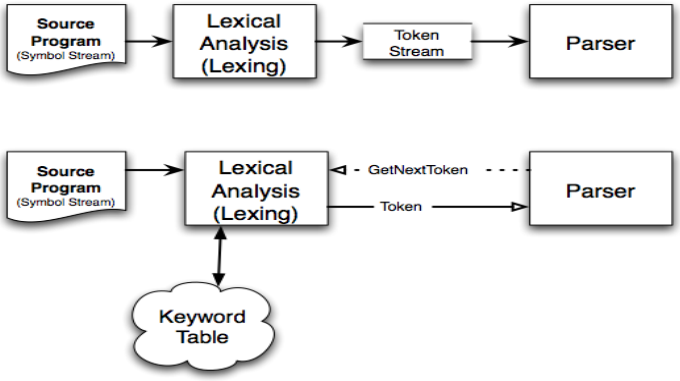
Semantic search systems consider various points including context of search, location, intent, variation of words, synonyms, generalized and specialized queries, concept matching and natural language queries to provide relevant search results.

Lexical analysis

Lexical analysis is the process of converting a sequence of characters into a sequence of tokens.



The purpose of lexical analysis is Transform a stream ofsymbols into a stream of tokens



LEXICAL ANALYZER

Scan Input

Identify Tokens

Create Symbol Table

Insert Tokens into table

Send Tokens to Parser

Token:

A classification for a common set of strings

Pattern:

The rules which characterize the set of strings for a token – integers [0-9]. Recall File and OS Wildcards ([A-Z]*.*)

Actual sequence of characters that matches pattern and is classified by a token Identifiers: x, count, name, etc.

Regular Expressions

A Regular Expression is a Set of Rules / Techniques for Constructing Sequences of Symbols (Strings) From an Alphabet.

Pattern extraction

For each entry, the definition is processed looking for words that are connected with the entry in Wikipedia by means of a hyperlink. If there is a relation in Word Net between the entry and any of those words, the context is analyzed and a pattern is extracted for that relation.

Pattern generalization

In this step, the patterns extracted in the previous step are compared with each other, and those that are found to be similar are automatically generalized.

CONCLUSION

I have proposed a semantic similarity measure using both page counts and snippets retrieved from a web search engine for two words. We proposed a lexical pattern extraction algorithm to extract numerous semantic relations that exist between two words. Moreover, a sequential pattern clustering algorithm was proposed to identify different lexical patterns that describe the same semantic relation. Both page counts-based co-occurrence measures and lexical pattern clusters were used to define features for a word pair.

FUTURE ENHANCEMNT

In future this application will be extended into all level of public orientation uses. For example this process will adapted to the health care domain, which will helps to get the valid treatments, medicines, and details about it. Not only it gives the result, it will produce more accuracy result with help of some efficient sites related to healthcare.

REFERENCES

[1] J. Hou and Y. Zhang, Effectively Finding Relevant Web Pages from Linkage Information, IEEE Transactions on Knowledge and Data Engineering, Vol. 15, No. 4, 2003.

[2] P Ravi Kumar, and Singh Ashutosh kumar, Web Structure Mining Exploring Hyperlinks and Algorithms for Information Retrieval, American Journal of applied sciences, 7 (6) 840-845 2010.

[3] M.G. da Gomes Jr. and Z. Gong, Web Structure Mining: An Introduction, Proceedings of the IEEE International Conference on Information Acquisition, 2005.

[4] R. Kosala, and H. Blockeel, Web Mining Research: A Survey, SIGKDD Explorations, Newsletter of the ACM Special Interest Group on Knowledge Discovery and Data Mining Vol. 2, No. 1 pp 1-15, 2000.

[5] S. Brin, and L. Page, The Anatomy of a Large Scale Hypertextual Web Search Engine,, Computer Network and ISDN Systems, Vol. 30, Issue 1-7, pp. 107-117, 1998.

[6] Wenpu Xing and Ali Ghorbani, Weighted PageRank Algorithm, Proceedings of the Second Annual Conference on Communication Networks and Services Research (CNSR '04), IEEE, 2004.

[7] J. Kleinberg, Authoritative Sources in a Hyper-Linked Environment, Journal of the ACM 46(5), pp. 604-632, 1999.

[8] C. Ding, X. He, P. Husbands, H. Zha, and H. Simon, Link analysis: Hubs and authorities on the world. Technical report: 47847, 2001.

[9] J. M. Klienberg, Authoritative sources in a hyperlinked environment. Journal of the ACM, 46(5):604-632, September 1999.

[10] S. Chakrabarti, B.Dom, D.Gibson, J. Kleinberg, R. Kumar, P. Raghavan,S. Rajagopalan, and A. Tomkins, Mining the Link Structure of the World Wide Web, IEEE Computer, Vol. 32, pp. 60-67, 1999.