SIMILARITY MEASURE USING LINK BASED APPROACH

[1]B. Bazeer Ahamed, [2]T.Ramkumar

[1]Department of Computer Science & Engg., [2]Department of Computer Applications
[1]Research Scholar, Sathyabama University [2]AVC Engg. College,Tamil Nadu, India
[1]bazeerahamed @gmail.com, [2]ramooad @yahoo.com

Abstract—Web search engines provide an efficient interface to vast information. This web search engine provides the most semantic relativity between the given words, and it will generate the semantic measures automatically, since data on the web is noisy, huge and dynamic. we propose and analyzed and visualized similarity relationships in Web data sets to identify how to integrate content and link analysis for approximating relevance.

Keywords- Related Pages, similarity, web retrieval, search, ranking

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.[1] 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.

Information Retrieval on the Web

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).[9]

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.

Ranked Retrieval Model

This feature makes Ranked Retrieval Model more

user- friendly than Boolean Retrieval Model and Extended Boolean Model[12]. 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[9].

2.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[3]. 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;[11] 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`.

1.Page Ranking

The purpose of Page Ranking is to measure the relative importance of the pages in the web[8]. There are many algorithms for this purpose.

The most important ones are: Hyper Search, Hyperlink-Induced Topic Search (HITS), PageRank, Trust Rank, and OPIC. 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.[10]

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, Page Rank 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.

Progress made in Hyperlink Retrieval

It is very important yet difficult to provide hyperlinks in a (HTML-) document. Hyperlinks dramatically improve content quality by presenting related work, contradictory positions, further information or simply by the continuation of the next page or by giving similar navigational information [6]. The question of how a web author can easily find such information remains, though.

Research on the area of hyperlinks has been carried out since the introduction of the World Wide Web service to the Internet. Kaindl et. al. present a compact history of the progress made so far [7].

Link retrieval research aims at generating hyperlinks if not completely automatically, at least with as little user interaction as possible. Very serious problems arise, though, when trying to retrieve hyperlinks of texts on a statistical base without any semantic knowledge. The results are of low quality [8]. Allan classified link types into three major groups: manual, automatic and pattern-matching [9]. The idea is to retrieve at least the easy-to-find links of the two latter groups and leave most of the former one to the user.

3.Rank similarity process

The most basic similarity measures are purely lexical. That is, they rely solely on matching the terms present

in the surface representations. Given two short segments of text, Q and C, treating Q as the query and C as the candidate we wish to measure the similarity of, we define the following lexical matching criteria:

1. Exact – Q and C are lexically equivalent.
2. Phrase – C is a substring of Q.
3. Subset – The terms in C are a subset of the terms in Q

For a specific query Q, let the set of documents returned by a standard search engine (e.g. VSR) be called the root set R.

- Initialize S to R.
- Add to S all pages pointed to by any page in R.

Add to S all pages that point to any page in R

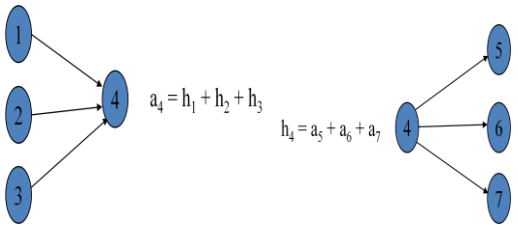


Fig 1. Link analysis measure

- Given a page, P, let R (the root set) be t (e.g. 200) pages that point to P.
- Grow a base set S from R.
- Run HITS on S.
- Return the best authorities in S as the best similar-pages for P.
- Finds authorities in the "link neighbor-hood" of P.

It is measured by the recall procedure in that the fraction of the documents those are relevant to the query that is successfully retrieved.

$$\text{Document} = \frac{|\{\text{non relevant document}\} \cap \{\text{retrieved document}\}|}{|\{\text{non-relevant document}\}|}$$

It is measured by

$$E = 1 - \frac{1}{\alpha / p + 1 - \alpha / r}$$

And their relationship is measured by

$F\beta = 1 - E$ where $\alpha = 1 / 1 + \beta^2$

Where P=number of pages, r= relevant pages, α = retrieved document , β=non-relevant document

To specify the attributes of hypertexts we chose the following settings

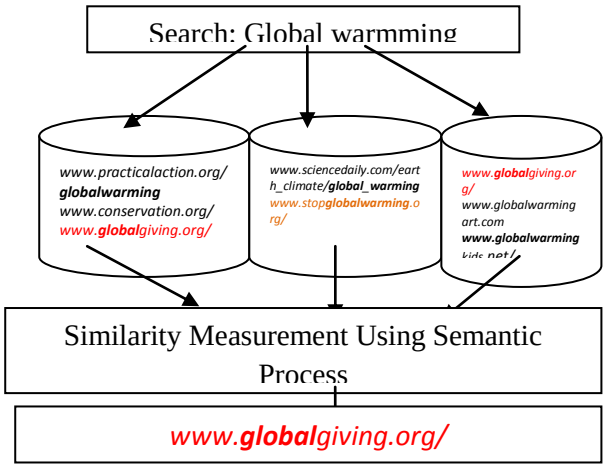
- Every important (weighted) keyword of the document is regarded as an attribute
- Every author of the document forms an attribute
- The creation date and expiration date of a document are subsumed to one attribute "validation"
- The publishing state and the version are combined to form the attribute "availability"
- The department information is one attribute "structure", but we make the restriction that each document must not belong to more than one department.

4.Experimental Result

Document is composed of many terms and important words are spread out documents. The importance of significance indicators assigned to the Web elements like title, heading, bold, anchor improves the ranking of the Web documents. Unlike text documents,[4]Web pages have certain characteristics such as structural information, hyperlinks and anchors which could serve as potential indicators of subject content. The relevance score of the document is assigned based on which term is matched and the part of the Web page in which the match is found. The annotation process of a web page is done with concepts of the ontology. Then, relations between individuals are discovered and instances are added. Document should be preprocessed to obtain semantic annotations and indexing of the document should be done. Examine the location of the annotated instance in the document. The annotation weights are calculated by combining the frequency and structures weights.Web search engines provide advanced features in that a user can specify how a query is matched with title of the page, text of the page, URL and links to the page, anywhere in the page.

Link Similarity

In information retrieval, we can assess a document-ranking system's effectiveness using precision-recall plots if the relevant set is known. Although evaluating how effectively we could rank Web pages based on, say, content or link similarity would be extremely interesting, it sets are unknown for the Web.[5] (Even large user studies can identify only subsets of relevant pages.) However, "query by example" retrieval systems, which use pages as queries, provide an interesting alternative. Following this approach, we can consider each page in the sample as an example by which to rank all other pages, and then use semantic similarity estimates to assess the rankings.



Finding the similarity measurement in different data sets, we obtained the given criteria by applying the formula,

Example 1:Q1: www.globalgiving.org/

C1: www.globalwarming.org

|{non relevant document }| = C1, |{retrieved document}|=Q1

Simplify (q1,C1)=||Q1(r1) ∩C1(r1) / Q1(r1)|| → 1

5/30=17%

Example 2:

Simplify (q2,C2)=||Q2(r2) ∩C2(r2) / Q2(r2)|| → 2

5/9=56%

Example 3:

Simplify (q3,C3)=||Q3(r3) ∩C3(r3) / Q3(r3)|| → 3

6/33=18.2%

equalize the equation

(q1,C1)=||Q1(r1) ∩C1(r1) / Q1(r1)|| ∪ (q2,C2)=||Q2(r2) ∩C2(r2) / Q2(r2)|| ∪ (q3,C3)=||Q3(r3) ∩C3(r3) / Q3(r3)|| Hence Avg(Q,C)=30.2

CONCLUTION AND FUTURE WORK

This work has presented an innovative technique for computing similarity among web pages. It has the capability of comparing a web page in larger data sets .Similarity among the pages contained in one particular set. These results can be used to rank pages as well as for finding relative page links. so far the algorithm use only in link analysis structure to compute similarity in future the tag analysis can be computed in composite score in similarity measurement.

REFERENCES

[1] Yuhua Li, Zuhair A. Bandar and David McLean, ―An Approach for Measuring Semantic Similarity between Words Using Multiple Information Sources‖,IEEE Transactions On Knowledge And Data Engineering,Vol.15,No.4,July/August200

[2] J. Gracia, R. Trillo, M. Espinoza, and E. Mena, ―Querying the web: A multiontology disambiguation method,‖ in Proc. of International Conference on Web Engineering, 2006, pp. 241–248.

[3] D. Bollegala, Y. Matsuo, and M. Ishizuka, ―Measuring semantic similarity between words using web search engines,‖ in Proc. of International Conference on World Wide Web, 2007, pp. 757–766.

[4] Elias Iosif and Alexandros Potamianos, ―Unsupervised Semantic Similarity Computation Between Terms Using Web Documents‖, IEEE Transactions On Knowledge And Data Engineering

[5] Lu Zhiqiang, Shao Werimin and Yu Zhenhua, ―Measuring Semantic Similarity between Words Using Wikipedia‖, 2009 International Conference on Web Information Systems and Mining

[6] F. J. Ricardo. Stalking the paratext: speculations on hypertext links as second order text. In Proceedings of the Ninth ACM Conference on Hypertext (Hypertext ’98), 1998. ACM

[7] H. Kaindl, S. Kramer. Semiautomatic generation of glossary links: A Practical Solution. In Proceedings of the Tenth ACM Conference on Hypertext (Hypertext ’99), 1999. ACM

[8] R. J. Glushko. Design issues for multi-document hypertexts. In Proceedings of the Second ACM Conference on Hypertext (Hypertext ’89), 1989. ACM

[9] J. Allan. Automatic hypertext link typing. In Proceedings of the Seventh ACM Conference on Hypertext (Hypertext '96), 1996. ACM

[10] S. Chakrabarti, B. Dom and P. Indyk, Enhanced hypertext categorization using hyperlinks, in: Proc. of the ACM SIGMOD International Conference on Management of Data, pp. 307–318, 1998.

[11] J. Kleinberg, Authoritative sources in a hyperlinked environment, in: Proc. of the 9th Annual ACM–SIAM Symposium on Discrete Algorithms, pp. 668–677, January 1998.

[12] T. Kistler and H. Marais, WebL — A programming language for the Web, in: Proc. of the 7th International World Wide Web Conference, pp. 259–270, Brisbane, Qld., April 1998.

❖❑❖