

How is link analysis efficient in providing relative documents to specified queries?

From all the papers you have read, which link analysis algorithm would you implement to your search engine project? Why? How would this algorithm be beneficial?

Explain how the HITS algorithm can be used to find different communities embedded in the world wide web.

Describe how link analysis can be used to generate more accurate results in a given search engine query.

The concepts of Data Mining are very important in search engines. In fact, the very idea of link analysis suggests some data mining. However, in most of the implementations which were discussed and reviewed, the data mining done was extracting information about a page based on the links alone. Implementations like these leave out the semantics of the words which make up each particular page which could also be extremely beneficial in the generation of topics and communities of web pages.

Will including the word semantics individual to each document improve the performance of some of the systems we have gone over?

Give reasons why each document's semantic content should or should not be looked at in determining link analysis?

Why was so much effort put into developing an efficient algorithm in the Trawling paper?

What are the differences between HITS and PageRank?

Give a basic description of Kleinberg's Algorithm (no equations, just written description) -- explain how it works, and how it categorizes the results.

How is link analysis used in search engines today? At what point in the "process" is it used (ie, crawler, parser, etc.)? How does it help the end user return useful results?

Survey the link analysis algorithms (e.g. HITS, PageRank, etc.). Explain the features and issues of each. Provide equations or pseudo-code. Explain variants/extensions of these algorithms. As mentioned in the papers, explain how these algorithms have been applied to applications besides in the assignment of ranks to search results.

Explain and differentiate the companion and cocitation algorithms. What are they intended to achieve? Which is easier to implement? On what previous work are these algorithms based on? How do these algorithms fare with respect to each other and with respect to other approaches? What service do the authors make use of in the implementation of these algorithms?

Briefly describe the HITS algorithm and explain its usefulness in developing algorithms based on link analysis for searching the web.

Explain the advantages and disadvantages of choosing to use a search engine based purely on link analysis as opposed to a search engine based purely on semantic analysis of web page text.

Describe the construction of root and base sets required by Kleinberg's HITS algorithm. Additionally, describe some of the drawbacks of that construction and specify some ways in which the sets might be improved.

Describe the concept of a "rank sink" with respect to the PageRank algorithm and its consequences for the ranking computation.

Characterize the random surfer model upon which the PageRank algorithm is based.

it has been noted in "Inferring Web Communities from Link Topology" that highly referenced commercial home pages appear with high relevance to a particular community when in fact they are not. How would you choose to alleviate this problem and why does it occur?

Why might a search for "spears" reveal the largest community of pages related to Britney Spears this year but "Spears of Destiny" or "Wooden spears" next year?

What are the basic assumptions of link analysis?

How to combine link analysis with text analysis in ranking and relevance function?