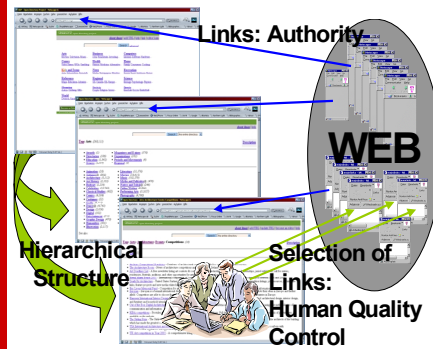


Abstract

This study analyzes the link behavior of web page authors and draws conclusions for the design of link analysis algorithms in web information retrieval. Back-links are more and more considered an important indicator for the authority of pages. A quantitative analysis of links to and from internet catalogues shows that the probability for a link to a page in a catalogue decreases drastically when the page is on a low level in the hierarchy. Furthermore, the number of links to a page in an internet catalogue does not correlate with the number of back links of the sites mentioned. As a consequence, link based authority measures need to be further refined in order to better reflect the cognitive processes involved in link creation. An mathematical model for the distribution is proposed.



Experiments

Our two experiments focus on the following issues arising in the context of web catalogues and their use:

- To which catalogue pages do web authors link? Are they more likely to point to a general high level entry page or do they value the work of the editing staff which may have selected high quality sites for their special topic?
- Does the analysis of in-links lead to consistent patterns for the authority of web pages? Do popular, highly linked catalogue pages also point to very authoritative sites?

Distribution of In-Links

The first experiment analyzed the relationship between the number of back-links of a catalogue page and the hierarchy level of that page. The information derived during the web mining process, shows a drastic decrease in the number of in-links for a decreasing hierarchy level. These findings show that humans are much more likely to set a link to a page which is positioned at a higher level of the hierarchy of a web site. This is surprising, since links are often based on topical similarity and therefore, more links to specific pages should be expected. Authors obviously prefer to create pages where visitors are expected to rely on browsing.

An additional crawl of non-catalogue pages confirms this finding, however, the distribution is less steep.

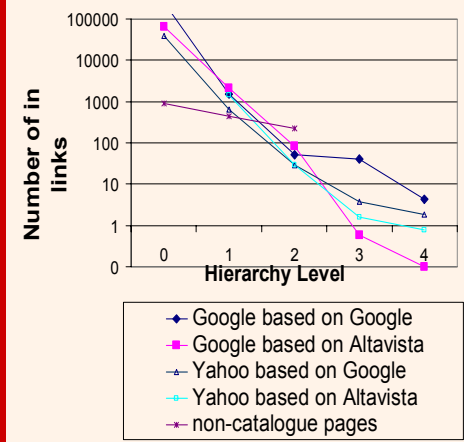
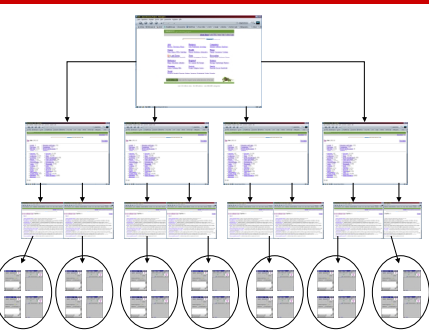


Fig.: Relationship between average number of in-links and hierarchy position



Approach

The reasons for humans to point to other pages in their own pages need to be further investigated, since quantitative link based measures are becoming an important factor in today's web search engines. This paper explores issues arising in the context of link behavior and draws conclusions for link-based authority measure in web information retrieval. The main findings are the following:

- The distribution of in-links for hierarchically organized web sites follows approximately an exponential distribution.
- The probability for an in-link greatly diminishes when a page is on a lower level of the sites' hierarchy.
- The calculated authority of a page in an internet catalogue does not correlate with the authority of the pages it refers to.
- The calculated authority of a page in an internet catalogue weakly correlates to the number of sub-categories in this page. No relations to other parameters were found.
- These results should be considered in the design of connectivity based web search algorithms.
- Currently, link analysis does not take into account the hierarchical position of a page. However, a page deep in the hierarchy which receives relatively many back-links deserves to be assigned a relatively high authority value.
- The authority solely calculated by link analysis does not reflect the complexity of human's evaluation of web pages. Further parameters need to be found.

Inconsistent Judgements

The second experiment intended to investigate the adequacy of the number of back-links as a quality indicator. If the number of back-links is a good indicator, a popular catalogue page should point to popular sites. The indicator should be consistent for catalogue pages and referred pages.

The analysis showed that there is no correlation between the two different counts of in-links. The number of in-links for a catalogue page is independent from the number of in-links to the entries, the sites that the page refers to. This finding confirms that there seems to be a rationale for separating between content quality (authority value) and referral quality (hub value).

Surprisingly, a significant correlation was found for another parameter, the number of sub categories. Whereas correlation with all other parameters remains close to zero, the number of sub categories exhibits a correlation over 0.5 for category level two.

	in-links based on Google	in-links based on Altavista	number of sub categories	number of entries
All pages	-0.05	-0.06	0.34	-0.13
Pages on level two	-0.10	-0.08	0.58	-0.13
Pages on level	-0.05	-0.06	0.04	0.10

Correlation between average number of in links for entry sites and four parameters of catalogues pages

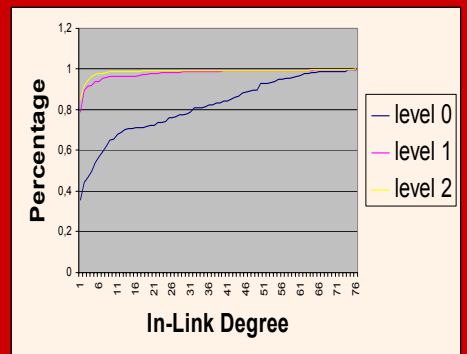


Fig.: Cumulative Link-Distribution for Non-Catalogue Pages

Distribution Model

The relation seen between average number of in-links hierarchy position seems to follow an exponential distribution. For catalogue pages, the lambda parameter is around 4.0 whereas for non-catalogue pages it is only around 0.3.

Outlook: Quality Retrieval

The hierarchical position of a page should be integrated as a factor for link analysis within search engines.

Quality retrieval needs to be based on more than one quality definition as it is the case in link analysis. Additional indicators for the quality of web pages need to be identified.