Information Science

Universität Hildesheim

Technical Report Nr.: 1/2003

Link Analysis and Site Structure: Refining Web Information Retrieval

Thomas Mandl
FB Informations- und Kommunikationswissenschaften
Universität Hildesheim – Hildesheim – Germany
mandl@uni-hildesheim.de

Abstract: Link analysis is the most important application of web structure mining and serves as a new knowledge source in web information retrieval. However, the mono-dimensional analysis of links neglects many other aspects. The results presented in this article show that the structure of a site affects the in-links for its pages. Link analysis algorithms need to be refined in order to account for that fact and the retrieval of high quality pages cannot solely be based on links as the only parameter.

1. Introduction

The Internet poses several new challenges for information retrieval and, at the same time, offers additional knowledge sources. One of the main problems for search engines is the heterogeneous quality of internet pages [HM02]. The common approach to automatically measure the quality of a page has been link analysis. The number of links pointing to a page are considered as main quality indicator. A large number of algorithms for link analysis has been developed [He00]. However, link analysis has several shortcomings.

The assignment of links is an inherently social process leading to strikingly clear overall patterns. The number of in-links for a web page follows a power law distribution [DK01]. In such a distribution, the median value is much lower than the average. That means, many pages have few in-links while few pages have an extremely high number of in-links. The authority value derived from the link-structure needs to be aggregate with the retrieval ranking. The power law distribution of PageRank values makes the value of some pages very dominant while other have little influence.

The influence of the power law is also visible for popularity measure in other social networks [Ba03]. It indicates that web page authors choose the web sites they link to without a thorough quality evaluation. Much rather, they act according to economic principles and invest as little time as possible for their selection. As a consequence, social actors in networks rely on the preferences of other actors. Pages with a high in-link degree are more likely to receive further in-links than other pages [PF02].

Another reason for setting links is thematic similarity. A co-citation matrix between web pages of different topics shows that most links are based on topical similarity [CJ02]. Quality decisions are not the only ones when setting a link.

Link analysis does not consider site structure. An analysis including the hierarchical position of a page within the site structure revealed that pages low in the hierarchy are more unlikely to receive in-links than top level pages like home pages [Ma02]. This analysis was based on internet catalogues. In this paper it is extended to a crawl of non catalogue pages. It seems that different parameters are necessary to describe the difference between catalogue and non catalogue pages. Further structural analysis hints that the hierarchical position has less influence on page specific features.

Large scale evaluation of web information retrieval can be done within the TREC (Text Retrieval Conference) initiative. TREC provides a testbed for information retrieval experiments. TREC is organized by the National Institute for Standards and Technology (NIST) which maintains a large real world collection of documents. Research groups apply their IR systems to this corpus and train it with the results from previous years. Their results can be NIST, where they are evaluated according to relevance. A few years ago, a web track has been introduced, where a large snapshot of the web is target of retrieval systems. In this context link based measures have been compared with standard retrieval systems on a large scale. The results show that the consideration of link structure does not lead to better retrieval performance for topical queries. Only for home page finding, improvement has been achieved by integrating the PageRank values into a retrieval algorithm [Ha01,GS01,Ya01,Ma03,]. As a consequence, more research and possibly refinements of link based algorithms are necessary.

2. Site Structure and In-Links

Web directories or internet catalogues are important services for the orientation in the internet. They usually intend to topically organize information sources and to introduce a certain level of quality control. Human editors monitor the web, evaluate and comment on pages. They order pages in a topical hierarchy.

One experiment presented in [Ma02] focused on the question to which catalogue pages web authors set their links. Are they more likely to point to a general high level entry page of a catalogue? Or do they value the work of the editing staff which may have selected high quality sites for their special topic? In the latter case, links to lower levels of the hierarchy would be more likely. The work presented in [Ma02] is extended to non-catalogue pages and page structure which is discussed in chapter 3.

2.1 Experiment and Results

The experiments are based on two German web catalogues. Pages in these web directories have been downloaded and analyzed. For this mining task, a document object model of each page has been generated. In addition, both Google and Altavista were queried about the number of in links to each of the pages in the internet catalogues. In the second experiment, we also queried these search engines for the number of back links for the entries contained in the catalogue pages. In total, some 3000 pages of the Google catalogue, some 4000 pages of the Yahoo directory and 8000 non catalogue pages from search engine results were downloaded and parsed.

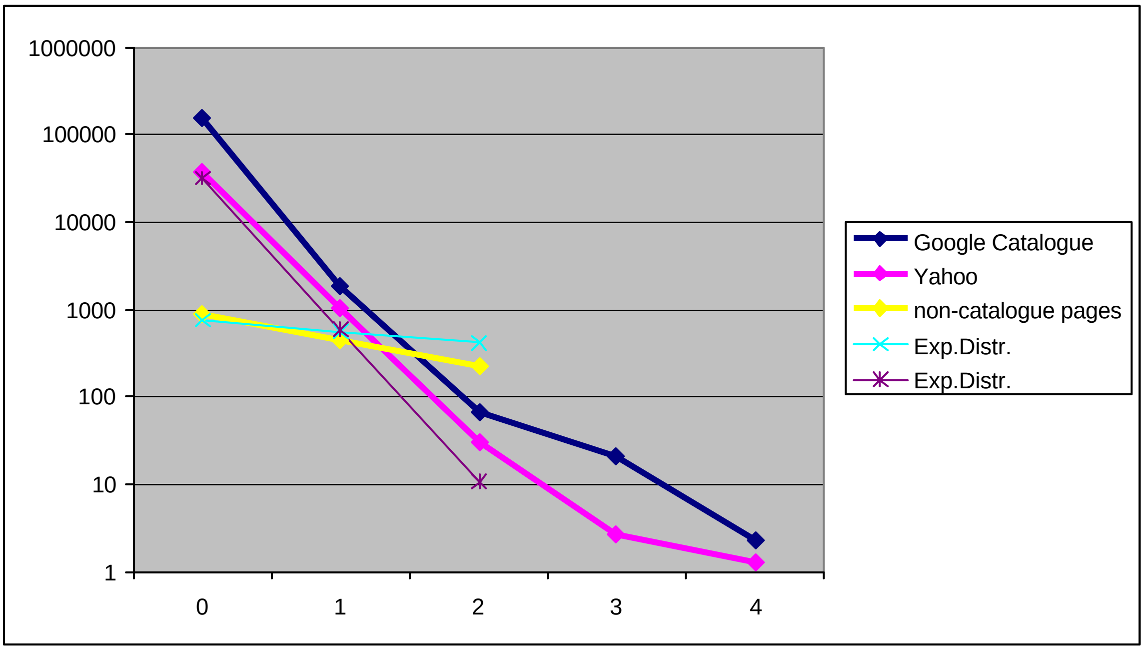

Figure 1: Relation between In-Links and Hierarchical Position of a Page

The information derived during the web mining process shows a drastic decrease in the number of links for a decreasing hierarchy level. The lower the pages are positioned in the site hierarchy, the less likely they are to receive back-links. The relationship between the average number of back-links for a page and its position in the hierarchy are shown in figure 1. There also seems to be a difference between the catalogue pages and non catalogue pages.

Assuming an exponential distribution, the relation for catalogue pages can be approximated by a parameter of 4 whereas for the non catalogue pages, a parameter of 0.3 gives a good approximation.

Another analysis of the non catalogue page crawl is shown in figure 2. The in-link degree is shown in relation to the percentage of pages with that in-link degree. It is clearly visible that the pages on different hierarchical levels exhibit a different cumulative distribution. Especially the pages on level 0 which are the home pages of the sites have a considerable higher probability of receiving many in-links than other pages.

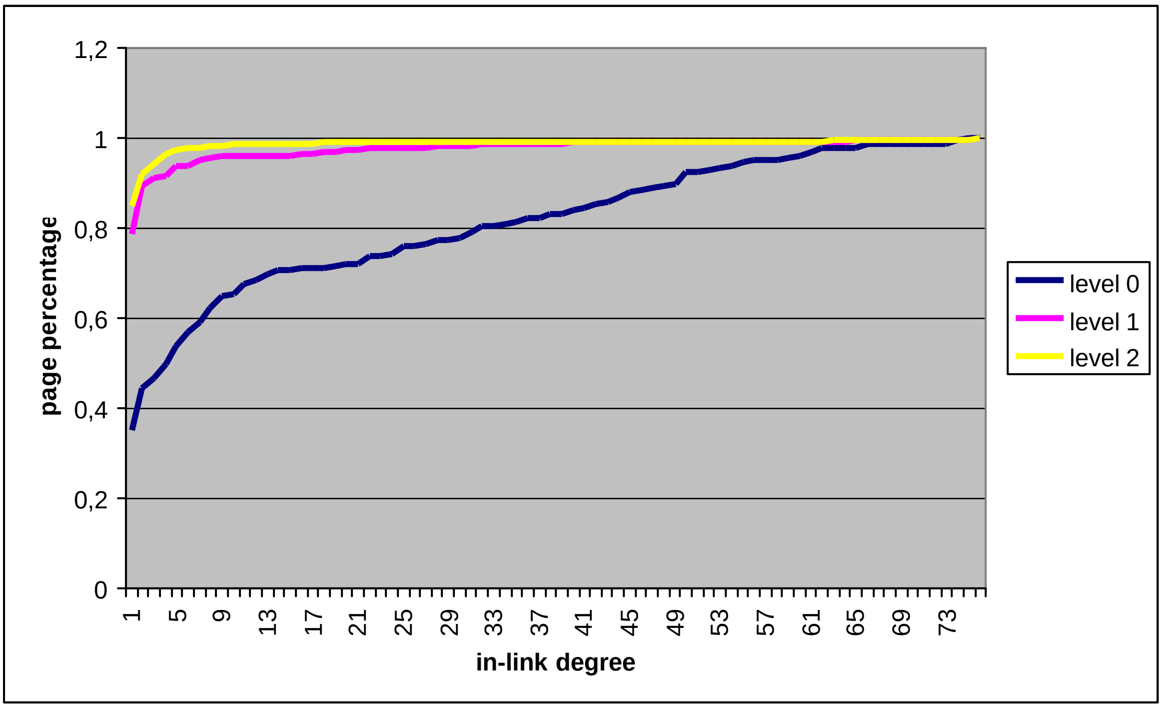

Figure 2: Cumulative Distribution for In-Links and Hierarchical Position of a Page

2.2 Consequences for link analysis algorithms

These results should be considered by the developers of web search engines. A cognitively more adequate exploitation of link structure may result in better retrieval results. Currently, link-based measures do 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 higherr authority value.

Furthermore, the analysis shows, how web page authors assume, that catalogue pages are best used. They prefer the presence of further options over pointing to a very specific page. Rather than targeting at pages with a narrow topical focus on lower hierarchical levels, they infer that users will benefit from further options for browsing. Web page authors stress the importance of navigation by browsing as information finding strategy in human computer interaction.

The experiments presented here also shed light on limits of the automatic evaluation of catalogue sites. This is especially interesting for systems which automatically search for the most authoritative catalogue page for a given topic. These would need to observe the hierarchical position of a topic within a catalogue.

3. Page Structure

Further structural analysis of the pages tried to reveal whether the hierarchical position has also effects on the structure of a page. If that is the case, e.g. home pages could be structurally very different from other pages and may therefore bias their accessibility by search engines. This may ultimately lead to higher in-link counts. In that case the results from the previous chapter may also have structural and technical reasons. However, this could not be verified. The structure of document is also subject to research within XML information retrieval [FG01].

The analysis of non catalogue pages was extended to page properties. We extracted the number of nodes in the document object model (DOM) derived from the page, the number of tables and the number of meta tags. The two first parameters are indicators for the structural complexity of a page whereas the number of meta tags indicates the extent to which the author aids indexing. These parameters were compared to both the in-link degree and the hierarchical position measured by the crawl order from the home pages.

A correlation analysis between both crawl order and in-link degree to the other parameters revealed no strong correlation. Only the number of tables and the number of DOM elements in a page correlated with a value of 0.24 and 0.21 respectively to the number of in-links.

A statistical analysis of the pages shows that pages on higher in the site structure tree tend to have more DOM elements with a lower deviation. They also seem to have more tables. However, these trends are very weak. The study needs to be repeated with a larger number of pages and sites.

Table 1: Statistics of the Analyzed Pages

	Average		
Hierarchy Level	**Nr DOM Elements**	**Nr MetaTags**	**Nr Tables**
0	295.2	8.7	13.6
1	292.8	11.9	10.9
2	252.5	9.1	9.5
	Standard Deviation		
Hierarchy Level	**Nr DOM Elements**	**Nr MetaTags**	**Nr Tables**
0	382.2	10.1	23.4
1	429.3	11.7	17.9
2	470.3	8.1	28.3

4. Outlook

Link analysis algorithms still suffer from many shortcomings. Some of them are presented in chapter 1. In addition, the analysis in [Ma02], revealed no correlation between the link based authority of a catalogue page and the link based authority of its entries. Therefore, a quality definition for web pages should not rely on one of one parameter only. It is necessary to find more reliable metrics for the quality of a web page which can be derived automatically.

References

[Ba03] Barabási, A.-L.: Linked: The New Science of Networks, Perseus, 2002.

[CJ02] Chakrabarti, S.; Joshi, M.; Punera, K.; Pennock, D.: The Structure of Broad Topics on the Web. In: Proc. Eleventh Intl World Wide Web Conf (WWW 2002). Honolulu, Hawaii. 7.-11.Mai. http://www2002.org/CDROMrefereed/338/

[DK01] Dill, S.; Kumar, R.; McCurley, K.; Rajagopalan, S.; Sivakumar, D.; Tomkins, A.: Self-Similarity in the web. In: Proc 27th Intl Conf on Very Large Databases (VLDB). 2001.

[FG01] Fuhr, N.; Großjohann, K.: XIRQL: A Query Language for Information Retrieval in XML Documents. In: (Croft, W.; Harper, D.;Kraft, D.; Zobel , J. eds.) Proc 24th Annual Intl Conf on Research and Development in Information Retrieval 2001. pp. 172-180.

[GS01] Gurrin, C.; Smeaton, A. (2001): Dublin City University Experiments in Connectivity Analysis for TREC-9. In (Voorhees, E.; Harman, D. eds.): The Ninth Text REtrieval Conf (TREC 9). 2001. http://trec.nist.gov/pubs/trec9/t9_proceedings.html

[Ha01] Hawking, D.: Overview of the TREC-9 Web Track. In (Voorhees, E.; Harman, D. eds.): The Ninth Text REtrieval Conf (TREC 9). 2001. http://trec.nist.gov/pubs/trec9/t9_proceedings.html

[He00] Henzinger, M.: Link Analysis in Web Information Retrieval. In: IEEE Data Engineering Bulletin, 23(3):3-8, 2000.

[HM02] Henzinger, M.; Motwani, R.; Silverstein, C.: Challenges in Web Search Engines. SIGIR Forum, 2002.

[Ma02] Mandl, T.: Evaluierung von Internet-Verzeichnisdiensten mit Methoden des Web-Mining. In (Hammwöhner, R.; Wolff, C; Womser-Hacker, C. eds.): Proc 8. Intl. Symposium für Informationswissenschaft. (ISI 2002). Regensburg. pp. 239-257.

[Ma03] Mandl, T.: Neuere Entwicklungen bei der Evaluierung von Information Retrieval Systemen: Web- und Multimedia-Dokumente. In: Information – Wissenschaft und Praxis vol. 54. 2003.

[PF02] Pennock, D.; Flake, G.; Lawrence, S.; Glover, E.; Giles, L.: Winners don't take all: Characterizing the competition for links on the web. In: Proc. National Academy of Sciences 99 (8) 2002. pp. 5207–5211

[Ya01] Yang, K.: Combining text- and link-based retrieval methods for Web IR. In (Voorhees, E.; Harman, D. eds.): The Ninth Text REtrieval Conf (TREC 9). 2001 http://trec.nist.gov/pubs/trec9/t9_proceedings.html