Estimating Content Quality in the World Wide Web

Johann Mitlöhner
Vienna University of Economics and Business Administration
Augasse 2–6, A-1090 Vienna, Austria
mitloehn@wu-wien.ac.at

Abstract: The World Wide Web is used by an increasing number of people as an ever-expanding source of information on almost every topic imaginable. However, useful data is often buried in large quantities of low-quality content. Estimation of content quality is valuable for diverse applications, such as search result ranking and direction of crawlers. In this text an approach is described to automatically determine author identity of web pages and compile data on author reputation in order to better estimate the quality of new content. The results of preliminary studies are presented which show the viability of the author reputation approach.

1 Introduction

Today many people are accustomed to using the Web as a source of information on all kinds of topics. Unfortunately, useful information is often burried among lots of irrelevant and low quality content. A lot of time is spent by people sifting through lots of useless material on the web.

We will define quality as usefulness, the ability to fulfill a certain purpose, such as providing desired information on a given subject. In most cases only human judgement can correctly state the quality of a page. Yahoo employs hundreds of editors to maintain a comprehensive web directory of quality sites on a large number of topics. The Open Directoy Project [She00] aims at identifying high quality web sites by using tens of thousands of unpaid human contributers. This approach works well in some areas; unfortunately, given the size and growth of the web there are many areas where that approach is infeasible. It is therefore desireable to estimate the quality of web content by other means. In this work author reputation is used to estimate content quality. It will be shown that using author reputation can result in an improvement in quality estimation in certain situations. The reputation approach is based on other methods of quality estimation which will be introduced in the next section.

2 Methods of Content Quality Estimation

One of the most well-known quality indicators is the in-degree $I(p)$ of a page p, defined as the number of pages that refer to p. Let L be the set of all links between pages, and $(p,q) \in L$ denote the fact that there exists a link from page p to page q, then $I(p) = \sum_{(q,p)\in L} 1$.

The idea underlying the in-degree count is that web page authors who include a link to other pages are implicitly or explicitly giving a recommendation for that other page. Obviously there are cases where this assumption is not valid. However, [ATH00] have used human expertise to compile ratings for a number of web pages on various topics and compared the results with automatically derived measures such as in-degree. The results show that in-degree is a good estimator of page quality. For each web page p we assume there exists an in-degree $I^*(p)$ describing its actual quality, which corresponds to the number of people interested in a particular subject who would find page p usefull enough to link to it if they knew page p. For several reasons cited below the actual in-degree $I(p)$ of most pages will generally be different from $I^*(p)$.

Note that in practical applications links coming from pages on the same host that p resides on are excluded from the in-degree count $I(p)$, since those links are likely to come from pages of the same author, and links between pages of the same author do not constitute a recommendation, but a necessary hypertext navigation element.

Other methods similar to in-degree have been derived from the link structure of the web, such as PageRank described in [PBMW98], and Authority/Hub values described in [Kle98].

PageRank is a recursive method which assigns ranks to pages based on the number of incoming links, weighted with the PageRank of the referring pages. The rank of some page p is $R(p) = c/n + (1-c)\sum_{(q,p)\in L} R(q)/O(q)$, where c is a constant with $0.1 < c < 0.2$, n is the total number of pages, and $O(q)$ is the number of outgoing links of page q. PageRank assigns high weights to incoming links that come from pages that have high PageRanks themselves, and that have few outgoing links. Starting with equal values for all pages in the set the rank of all pages is calculated repeatedly, and generally converges after several dozens of iterations. The popular Google search engine (www.google.com) uses this approach to rank its query results.

The HITS algorithm uses a similar idea but distinguishes between two page attributes: containing information themselves (authorities), and pointing to other authorities (hubs). Starting with $H(p) = 1$ and $A(p) = 1$ for all pages p iterate through $A(p) = \sum_{(q,p)\in L} H(q)$ and $H(p) = \sum_{(p,q)\in L} A(q)$, normalizing the A and H vectors after each step. Again, the A and H values generally converge quickly.

In order to calculate in-degrees a database of pages and links has to be maintained. Web crawlers are used to accumulate that data.

3 Focused Crawling and Web Dynamics

Many users apply search engines to find pages which contain a given set of keywords. These engines use web crawlers to maintain their database. A crawler visits pages, reports data such as document title, keywords, and outgoing links to the database, and places the outgoing links in a queue for subsequent crawling. Given the enormous size and fast rate of growth of the web, crawlers face a variety of problems. It is increasingly difficult to crawl more than a fraction of the web. Times between repeated visits to individual pages can reach several months for major search engines [LG99]. Quality estimation based on in-degree and keyword-based retrieval can be tricked by page authors and often produces undesirable results.

Focused crawlers [CvdBD99] aim at avoiding some of those problems. A focused crawler limits its visits to pages which fulfill certain criteria. A start set of documents on a given topic can be used to guide the progress of the crawler. Each new page is compared to the pages in the start set. Document similarity measures such as Cosine or Jacquard can be used to calculate the average degree of similarity s of the new page with all pages in the start set. The crawler only proceeds processing that page if s is above some predefined limit l. Such a crawler can get stuck with only a small fraction of the relevant pages when the similarity limit l is set too high; however, with proper settings it may quickly gather a significant fraction of the pages on a given topic. Since that set is still much smaller than the whole web the crawler can frequently re-visit all pages in the set and monitor changes, such as new links and newly created pages.

Web pages are created, changed, and deleted all the time. Unfocused crawling takes much time, and during that time the web continues to change, which means that crawlers cannot report the actual state of even a small fraction of the web at any given point in time. Since changes in the web mean changes in link structure a time argument is added to the in-degree $I(p,t)$, denoting the number of pages that link to p at time t. Note that at t_0 when p is created $I(p,t_0)$ will be very small and often far from $I^*(p)$. Quality estimation by in-degree works well for pages that have existed for some time, since it is necessary for other people to find a new page and read and evaluate its content before they possibly add links to that page in their own web site. Since many people use search engines to discover new pages, the search engine must first visit and index newly created pages before users can find it. Given the long re-visit intervals for major search engines stated above several months may pass until the in-degree $I(p,t)$ reaches the level $I^*(p)$ appropriate to the actual quality of page p.

This means that at any point in time the significant fraction of new content is not at its I^* level. In-degree is not a good quality estimator for those pages. This situation is all the more unsatisfactory since the newly created pages are likely to contain up-to-date information. In the next section an approach is examined which tackles this problem.

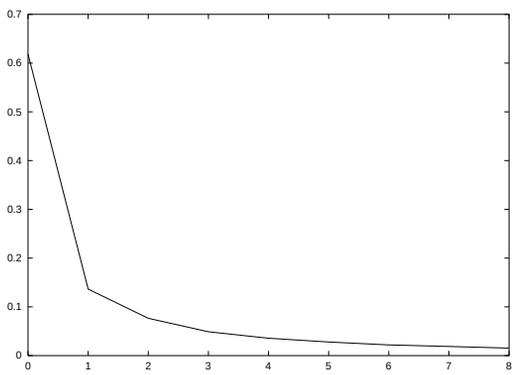

Figure 1: Relative frequency of in-degree for pages with author information

4 Improved Quality Estimation by Author Identification

In the same way that brand names offer some information about the quality of newly introduced products the identity of page authors might offer information on the quality of newly created pages.

A random crawler has been developed [Mit01] which models the behaviour of a random surfer as described in [HHMN00]. The random crawler starts with a set of seed URLs and randomly picks one URL to begin with. From that page it follows one of the outgoing links with probability $(1 - c)$, or it picks one URL from the set of already known URLs (including the seed set) with propability c. For pages that contained information about the page creator (see below) the crawler requested the in-degree for that page from a major search engine (www.alltheweb.com), excluding links coming from the same host the page resided on. Fig. 1 shows the relative frequencies of in-degrees for the 1.4 million pages crawled.

The crawler analysed the current page and tried to identify the author. At first glance it seems almost impossible to automatically identify the author of a page; however, it turned out that 28% of the pages contained a single MAILTO tag, a further 9% contained a single mail address somewhere in the page, and another 9% contained a META AUTHOR tag in the header. The address or the name given were assumed to identify the author. There are situations where this approach is problematic, such as email addresses of the form office@somewhere.org, where in fact several people are responsible for page creation. For reasons given below these problems affect the magnitude of the results described below, but not the general direction.

If author identification is to be useful in estimating the quality of newly created pages some individual quality level $Q(a)$ for a given author a must exist. Imagine the following experiment: for each author who created at least two pages we set up a basket and put all pages from that author into the basket. Now we pick two pages from each basket. If there was no individual quality level the information about the in-degree of one page of any

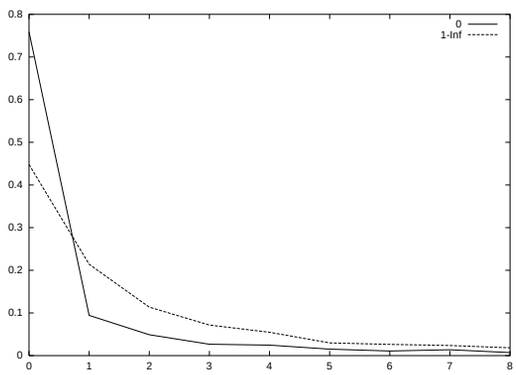

Figure 2: Relative frequency of second pick in-degree for authors identified by META tag, for first pick in-degree equal zero, and greater zero

author would provide no clue to the in-degree of the other page. Fig. 2 shows the in-degree distribution of the second page picked for two cases: (1) the first page had in-degree equal to zero, and (2) the first page had in-degree greater than zero. The distributions are clearly different. Other partitions and chi-square testing on the data plotted in fig. 2 prove this impression:

	$[1,4)$	$[4,20)$
$[1,4]$	254	108
$[4,20)$	159	164

In this setting of low/medium in-degree for first page (row) and second page (column) there is still a higly significant distribution difference. There is a significant dependency of the second-pick in-degree on the first; however, the dependency is very weak. Nevertheless, these results indicate that author information can indeed provide some clue to the actual quality of newly created web content. The average in-degree of all older pages from author a can serve as a reputation value. Newly created pages are better estimated by this value than by its current in-degree, which is necessarily near zero. We therefore need a way to automatically identify newly created pages, since only these pages are candidates for quality estimation by author reputation. Older pages are already near their I^* level.

It is difficult to automatically determine the age of a given page without knowing its context. Information from the web server such as the last-modified date is not useful, since the amount of modification is unknown. The only reliable method of determining the age of pages is to repeatedly crawl the whole set, spotting new pages by comparing with previous crawls. This approach is infeasible for the whole web, for reasons of size and growth given above. However, focused crawlers can be used for selected subsets. Each crawler can keep track of a moderate collection of pages, identifying newly created material und monitoring in-degree dynamics for all pages.

A focused crawler is being developed to continually monitor small subsets of the web. The Open Directory Project provides a taxonomy and start sets for use in crawling and

measuring document similarity. The quality rankings of the pages found resulting from combined in-degree and author reputation will be used to provide more up-to-date and more complete directories on selected topics.

It is expected that the author reputation approach will result in better quality estimation in certain situations. These situations are characterised by a sufficiently high content dynamic and a sufficiently stable user community. In a static setting where very few pages are created the indegree counts of most pages already correspond to their actual quality, rendering the reputation approach nearly useless. In a community with lots of entries and exits data on author reputation cannot be compiled to a sufficient degree. Therefore, some topics are expected to show improvement in page quality estimation, while others are unsuited for this approach.

The currently applied methods of automatic author identification show room for improvement. Email addresses of the form office@somewhere.org introduce noise that narrows the gap between the two in-degree distributions in fig. 2. More elaborate page text pattern processing is expected to result in improved author identification.

5 Conclusion

In this work quality estimation methods have been described that have been shown to correlate to human quality judgement. These methods are based on the link structure of the web. With the highly dynamic nature of the world wide web that structure is continually changing, and human-edited directories as well as general-purpose search engines meet with increasing difficulties in keeping up with the growth of the web. Focused crawlers can cope with the comparatively small number of pages belonging to selected topics and can automatically identify newly created content, since it is possible to repeatedly visit a small set of pages and spot newly created ones.

The quality of new content cannot adequately be estimated with in-degree based methods. It has been shown that authors maintain individual quality levels, and preliminary results indicate that author information can be used to better estimate the quality of new pages. With the large amount of new web content that is being created every day every improvement in better estimating the quality of that new content means less time for people browsing useless pages.

The approach suggested here uses author reputation to better estimate new content. There are drawbacks to this approach: obviously, a well-known author does not always guarantee for good web pages. However, indegree-based estimation suffers from a similar problem, since links do not always constitute a 'recommendation'. Many links should not be understood as a statement about content quality. Links are made for a variety of other reasons, such as commercial and organisational ones, among others.

One could also argue that 'newcomers' are at a disadvantage. No reputation is available for new authors, which means that newly created content will stay at its low initial indegree count. However, the situation is similar with solely indegree-based estimation. Since both approaches rely on the judgement of the user community interested in a particular topic

sufficient time must pass to allow for human quality judgement.

References

[ATH00] B. Amento, L. Terveen, and W. Hill. Does Authority Mean Quality? Predicting Expert Quality Ratings of Web Documents. In *Proc. SIGIR'2000 Conf. Research and Development in Information Retrieval*, New York, 2000. ACM Press.

[CvdBD99] Soumen Chakrabarti, Martin van den Berg, and Byron Dom. Focused crawling: a new approach to topic-specific Web resource discovery. *Computer Networks (Amsterdam, Netherlands: 1999)*, 31(11–16):1623–1640, 1999.

[HHMN00] Monika R. Henzinger, Allan Heydon, Michael Mitzenmacher, and Marc Najork. On Near-Uniform URL Sampling. In *Proceedings of the Ninth International World Wide Web Conference*, Amsterdam, 2000. Elsevier Science. http://www.www9.org/w9cdrom/88/88.html.

[Kle98] J. Kleinberg. Authoritative Sources in a Hyperlinked Environment. In *Proc. of the Ninth Ann. ACM-SIAM Symp. Discrete Algorithms*, New York, 1998. ACM Press.

[LG99] Steve Lawrence and C. Lee Giles. Accessibility of Information on the Web. *Nature*, 400:107–109, 1999.

[Mit01] Johann Mitlöhner. Web Content Quality and Author Reputation. In *Proc. of the ICEC 2001 Int. Conf. on Electronic Commerce*, Vienna, 2001. http://icec.net/icec2001.

[PBMW98] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The PageRank Citation Ranking: Bringing Order to the Web, 1998. Stanford Univ. Working Paper 1999-0120, http://www-db.stanford.edu/backrub/pageranksub.ps.

[She00] Chris Sherman. Humans Do It Better: Inside the Open Directory Project. *ONLINE*, July 2000. http://www.onlineinc.com/onlinemag/OL2000/sherman7.html.