Does "Authority" Mean Quality?
Predicting Expert Quality Ratings of Web Documents

Brian Amento[1], Loren Terveen, and Will Hill
AT&T Shannon Laboratories
180 Park Avenue
Florham Park, NJ 07932 USA
{brian, terveen, willhill}@research.att.com

KEYWORDS
exploiting hyperlink structure

ABSTRACT
For many topics, the World Wide Web contains hundreds or thousands of relevant documents of widely varying quality. Users face a daunting challenge in identifying a small subset of documents worthy of their attention.
Link analysis algorithms have received much interest recently, in large part for their potential to identify high quality items. We report here on an experimental evaluation of this potential.
We evaluated a number of link and content-based algorithms using a dataset of web documents rated for quality by human topic experts. Link-based metrics did a good job of picking out high-quality items. Precision at 5 is about 0.75, and precision at 10 is about 0.55; this is in a dataset where 0.32 of all documents were of high quality. Surprisingly, a simple content-based metric performed nearly as well; ranking documents by the total number of pages on their containing site.

1. INTRODUCTION: THE PROBLEM OF QUALITY

Finding documents on the World Wide Web *relevant* to a given interest typically is easy. Suppose you're interested in a Television show like The Simpsons. Search engines like Google or AltaVista return tens of thousands of items, and even human-maintained directories like Yahoo or UltimateTV contain dozens to hundreds of items.
However, these items vary widely in *quality*, ranging from large, well-maintained sites to smaller sites that contain specialized content to nearly content-free, completely worthless sites. No one has the time to wade through more than a handful of items.
The quality of a web site inherently is a matter of human judgement. Major factors influencing quality judgements include site organization and layout, as well as the quantity and uniqueness of information. Note that even small sites may be judged high-quality if they cover a particular sub-area well.
We treat quality and relevance as distinct notions, rather than viewing quality as just an aspect of relevance judgements. Perhaps an example can clarify the distinction. It seems natural to view a student paper and a collection of literary criticism as equally relevant to their topic, e.g., Shakespeare's sonnets, while allowing that the latter is of much higher quality.
Link analysis algorithms have received much attention recently, in large part for their potential to help with this problem. The basic intuition is that a hyperlink from document A to document B implies that the author of document A thinks document B contains worthwhile information. Thus, counting the links to a document may yield an estimate of the document's quality. More sophisticated algorithms have been developed that build on this intuition.
However, there has been little empirical evaluation of these algorithms. This leaves a fundamental issue unresolved – do link-based metrics work, i.e., do they correlate with human judgements of quality? We're actually interested in a more general question, namely whether *any* metrics we can compute for web documents are good predictors of document quality. Accordingly, we'll investigate content-based as well as link-based metrics.
We encountered several other questions while investigating this issue. First, we wondered to what extent topic experts agree on the quality of items within a topic. If human judgements vary widely, this suggests limits on the utility of automatic methods (or perhaps that collaborative filtering, which can personalize recommendations for an individual, may be more appropriate). More fundamentally, it would call into question whether a shared notion of quality even exists. Conversely, if experts do tend to agree in their quality judgements, our confidence in the concept of quality will be bolstered, even if it is difficult to give a precise definition.

[1] Also with Department of Computer Science, Virginia Tech.

Second, we wondered whether there were any significant differences between various link analysis algorithms – for example, would one score documents D_1, D_2, and D_3 highly, while another scored D_4, D_5, and D_6 more highly? If there are no such differences, then an algorithm can be chosen for other factors, such as efficiency.

The remainder of the paper is organized as follows. First, we outline our research program, to clarify the context for this experiment, and compare it to related efforts. Next, we describe a large study we carried out and explain how we obtained the dataset we analyzed for this paper. The heart of the paper consists of a description of the analyses we did to answer our three research questions and an interpretation of the results we obtained. We close by suggesting areas for future work.

2. PREVIOUS AND RELATED WORK

Our research program investigates the major information-finding problems that users of the world wide web face:

- finding collections of items *relevant* to their interests; we focus on the case where users are interested in fairly general *topics*, such as a television show, musical artist, or health concern, rather than a specific *query*;
- identifying *high-quality* items within a collection;
- finding items that contain a certain *category* of information, e.g., episode guides (for a television show) or song lyrics (for a musician);
- creating and maintaining personalized subsets of items; users create such collections for their own personal use as well as for sharing with others.

We have addressed these problems by developing algorithms, implementing them in web crawling and analysis tools, and creating interfaces to support users in exploring, comprehending, and personalizing collections of web documents [1, 2, 13]. We have moved from the web *page* (URL) to the web *site* as the basic unit of interaction and analysis. We discuss the notion of a site briefly since the experiment we report on here used this concept. A site (multimedia document) is an organized collection of pages on a specific topic maintained by a single person or group. Sites have structure, with pages that play certain roles (front-door, table-of-contents, index). A site is not the same thing as a domain: for example, thousands of sites are hosted on www.geocities.com..

Our webcrawler/analyzer heuristically groups the URLs it fetches into sites by examining the URL strings. The basic intuition is:

- if URL A is a prefix of URL B, then assume that A and B belong to the same site, A is the root page of the site, and B is an internal page.

In practice, this simple rule is augmented with both host-specific heuristics (e.g., for hosts like geocities or tripod) and general heuristics (e.g., to select a root page when two or more URLs are at the same depth).

When we aggregate URLs into sites, we aggregate links too: we record a link from site A to site B if any URL contained on site A links to any URL contained on site B.

Much recent research has focused on collections of hyper-linked documents, specifically the World Wide Web. Several systems have explored interaction techniques to help users explore and comprehend collections of items. SenseMaker [1] focuses on supporting users in the contextual evolution of their interest in a topic. It attempts to make it easy to evolve a collection, e.g., expanding it through query-by-example or limiting it by applying a filter. Scatter/Gather [11] supports the browsing of large text collections, allowing users to iteratively reveal topic structure and locate desirable documents. WebBook and WebForager [6] allow users to define, visualize, and manipulate groups of related web pages.

More relevant to the concerns of this paper are techniques that analyze link structure to rank and group items. Pitkow and Pirolli developed clustering algorithms based on co-citation analysis [12] and categorization algorithms that utilized hyperlink structure [10].

Kleinberg formalized the quality of documents within a hyper-linked collection using the concept of *authority* [8]. At first pass, an authoritative document is one that many other documents link to. However, this notion can be strengthened by observing that links from all documents aren't equally valuable – some documents are better *hubs* for a given topic. Hubs and authorities stand in a mutually reinforcing relationship: a good authority is a document that is linked to by many good hubs, and a good hub is a document that links to many authorities. Kleinberg developed an iterative algorithm for computing authorities and hubs. He presented examples that suggested the algorithm could help to filter out irrelevant or poor quality documents (i.e., they would have low authority scores) and identify high-quality documents (they would have high authority scores).

Several researchers have extended this basic algorithm. Chakrabarti et al [7] weight links based on the similarity of the text that surrounded the hyperlink in the source document to the query that defined the topic. Bharat & Henzinger [4] made several important extensions. First, they weighted documents based on their similarity to the query topic. Second, they count only links between documents from different *hosts*, and average the contribution of links

from any given host to a specific document. That is, if there are k link from documents on one host to a document D on another host, then each of the links is assigned a weight of 1/k when the authority score of D is computed. In experiments, they showed that their extensions led to significant improvements over the basic authority algorithm.

The notion of a "host" is similar to our "site". However, Bharat & Henzinger do not specify the exact definition of a host. If hosts are taken to be domains, critical information can be lost. For example, in one of the popular entertainment topics we included in our study, the television show Buffy The Vampire Slayer, 134 of the 258 distinct sites were hosted on geocities.com. If all these sites were considered to belong to the same host, a majority of interesting links would have been ignored.

PageRank [9] is another link-based algorithm for ranking documents. Like Kleinberg's algorithm, this is an iterative algorithm that computes a document's score based on the scores of documents that link to it.

To summarize, much recent research has experimented with link-based algorithms. A major motivation for these algorithms is that they can be used to compute measures of document quality. Yet there is little empirical evidence that what these algorithms compute (site in-links, authority scores, PageRank scores) actually correlates with human quality judgements.

3. EXPERIMENT

We recently carried out a large-scale empirical investigation of how web users look for quality items and what sort of support system could help users with this task.

We selected 5 popular entertainment topics for the study, the television shows Babylon 5, Buffy The Vampire Slayer, and The Simpsons, and the musicians Tori Amos and the Smashing Pumpkins. Popular entertainment is one of the main interests people follow on the web. A study of 1.1 million queries issued to the Magellan search engine between March 1997 and April 2000 supports this claim. We found that 42% of the queries were about popular entertainment. Its popularity alone makes this domain worthy of investigation. Further, we believe that it is similar to other domains characterized by rich content and many links between sites, including popular scientific topics such as Mars exploration.

3.1 Datasets

For the purposes of this study, we wanted to begin with a set of relevant web documents for each topic. We thus decided to use a web directory, where humans categorize URLs by topic. Yahoo is the most popular general purpose web directory, so we used it to obtain 5 sets of web sites. Our examination of these sets of sites show that they vary widely in quality, but nearly all are directly on topic.

The first phase of the experiment was a user study. We recruited 40 subjects from a local university. Their task was to select the 15 best items for a topic (subjects were randomly assigned a topic and interface). We defined the "best" items as those that together gave a useful and comprehensive overview for someone wanting to learn about the topic.

Subjects used either the Yahoo interface or our research prototype [1, 2] to explore, browse, and select items. In related work, we are analyzing in detail how subjects used the two interfaces [1]. However, the subjects' results play only a single role here. We needed to obtain a set of high-quality items to compare various algorithms to, and used items that the subjects selected as an initial version of this set. The intuition is that high-quality items are very likely to be selected by at least one subject.

In a second phase of the experiment, topic experts rated the quality of the sets of items obtained from the subjects. We solicited self-identified topic experts from AT&T and Virginia Tech, offering each $20 for participation. We obtained 4 experts for The Simpsons, and 3 for each of the other topics. Experts rated the quality of items on a scale of 1 (worst) to 7 (best). Experts rated items by filling out a web-based form; the form presented no information about items other than the URL, so experts had to browse the sites to judge their quality. Each expert's form presented items in a random order.

3.2 Url and Site Graphs, Url and Site Features

To compute link and content based metrics to compare with the expert ratings, we had to analyze the web neighborhood surrounding the items. We did this by applying our webcrawler/analyzer to each collection of items we obtained from Yahoo.

Starting from these seeds, the crawler constructs the surrounding neighborhood. Link and text similarity heuristics are used to select URLs to fetch and add to the neighborhood. In addition, for the purposes of this experiment, we limited the crawler to consider only urls on the same site as one of the seeds; we did this by accepting only URLs which contained some seed URL as a prefix.

When the crawling is complete, URLs are aggregated into sites (as described above). In addition to the basic URL graph – whose nodes are URLs, and whose edges represent hyperlinks between URLs – this results in a site graph – whose nodes are sites, and whose edges represent a hyperlink from (any url on) one site to (any url on) another.

From these graphs, we computed 5 link-based features; in and out degree, Kleinberg's authority and hub scores, and the PageRank score. In all cases, we computed features for both the site and the root URL of the site. Computing these metrics at the site level was straightforward. When we computed at the url level, we followed Bharat & Henzinger [4] by (1) counting only links between urls on different sites, and (2) averaging the contribution of links from all the URLs on one site to a document on another.

The crawler also computes a set of content-based features for each url. Page size and the number of images and audio files are recorded. This information is aggregated to the site level, and the total number of pages contained on each site also is recorded.

Finally, the crawler computes text similarity scores. Although we consider relevance and quality to be different notions, we wanted to test whether relevance would help predict quality. The crawler uses Smart [5] to generate a centroid – a weighted vector of keywords – from the content of the seed items for each topic. The relevance score of each item is based on the inner product similarity of the item's text to the centroid. And for each site, the relevance score of the root page, the maximum relevance score of any contained page, and the average relevance scores of all contained pages are recorded.

Each of the features induces a ranking of the items in our dataset. In subsequent analysis, we examine how well the various rankings match human quality judgements. To summarize, here is a list of all the features we used:

- In degree – number of sites that link to this site,
- Kleinberg's Authority Score,
- PageRank Score – link-based score used in Google [2],
- Out degree – number of sites this site links to,
- Kleinberg's Hub Score,
- Text relevance score – similarity to topic seed text,
- Size (# of bytes and # of contained pages),
- # of images, and
- # of audio files.

4. RESULTS

4.1 Do experts agree?

We first investigated how much experts agreed in their quality judgements. To the extent they do agree, we gain confidence that there is a shared notion of quality within the topic areas we investigated. We did two computations to measure agreement. First, we correlated the scores assigned to items by each pair of experts for each topic. (Recall that we had 4 experts for The Simpsons and 3 for all other topics.) We used the Pearson product-moment correlation since the expert averages represent interval data, ranging from 1 to 7. Table 1 presents the results. It shows that almost all pairs of experts were highly correlated in their judgements of item quality (all correlations were significant, $p<0.01$).

	Correlations between pairs of experts						
Topic	**1-2**	**2-3**	**1-3**	**1-4**	**2-4**	**3-4**	**Avg**
Babylon 5	0.91	0.92	0.76				**0.87**
Buffy	0.75	0.79	0.83				**0.79**
Smashing Pumpkins	0.80	0.73	0.69				**0.74**
Tori Amos	0.61	0.63	0.50				**0.58**
Simpsons	0.52	0.59	0.50	0.75	0.59	0.59	**0.59**
Total							**0.71**

Table 1: Expert agreement, using correlation

We did a second analysis that abstracted the expert judgements a bit. Rather than using the exact scores that experts assigned to items, we categorized each item into one of two bins – "good" items were those that an expert rated 5, 6, or 7, and "other" items were all the rest. (We use this categorical notion of quality in many of the remaining analyses.) For each topic, we computed the set of items that all experts assigned to the same category, as well as pairwise agreement (shown as Avg PW Agr in Table 2) between each pair of experts.

Table 2 presents the results, which are quite similar to the correlations presented above. On average, across topics, all experts agreed on the category for 65% of items. Pairs of experts agreed 78% of the time.

Topic	**# items**	**#Agr**	**%Agr**	**Avg PW #Agr**	**Avg PW %Agr**
Babylon 5	40	31	0.78	34.0	0.85
Buffy	41	28	0.68	32.3	0.79
Simpsons	39	24	0.62	30.7	0.79
Smashing Pumpkins	41	28	0.68	32.3	0.79
Tori Amos	42	21	0.50	28.0	0.67
Average	40.6	26.4	**0.65**	31.5	**0.78**

Table 2: Expert agreement, using categories

These results suggest that experts agree on the nature of quality within a topic, and that the expert judgements thus can be used to evaluate rankings obtained by algorithms. However, there is some variation between topics; Babylon 5 experts agreed the most, Tori Amos experts the least, and the other three topics were in the middle. Some lack of agreement may be due to properties of the topics. For example, we noticed that one or two Tori Amos sites were of quite high quality, but somewhat tangential relevance to the topic. Some experts' quality judgements may be influenced by the relevance. Second, some variation in opinions is inevitable, particularly in the area of popular entertainment,

where there is no objective quality standard. One expert may be more interested in one type of content than another (e.g., song lyrics vs. tour schedules). Some experts may have highly idiosyncratic tastes. Where tastes do differ significantly, a collaborative filtering approach ultimately may be necessary. To get the best information for *you*, you may have to inform the system about your preferences, so it can find experts with similar preferences, and recommend items that they like.

4.2 Are different link-based metrics different?

The second issue we investigated was whether the three link-based metrics – in degree, authority, and PageRank – ranked items differently.

Since the different metrics use different scales that do not maintain a linear relationship, we converted raw scores into ranks and used Spearman's rho rank correlation on the resulting ordinal data. We computed correlations between each pair of metrics.

Topic	In/Auth	In/PR	Auth/PR
Babylon 5	0.97	0.93	0.90
Buffy	0.92	0.85	0.70
Simpsons	0.97	0.99	0.95
Smashing Pumpkins	0.95	0.98	0.92
Tori Amos	0.97	0.92	0.88
Average (Spearman)	**0.96**	**0.93**	**0.87**
Average (Kendall)	**0.86**	**0.83**	**0.75**

Table 3: Metric similarity, using correlations

Table 3 presents the results. The correlations were extremely high (and were all significant, $p<0.01$). We also computed the Kendall tau rank correlation. Correlations again were high, although not quite as high as Spearman's rho; the final row in Table 3 presents the average Kendall correlations.

Second, we computed the intersections between the top 5 and top 10 items as ranked by the three metrics. Table 4 presents the results. Again, there is great agreement. For example, in-degree and authority have an average intersection of 8.4 of the top 10 items, and all three metrics agree on an average of 6.4 of the top 10 items.

Topic	I/A 5	I/P 5	A/P 5	All 5	I/A 10	I/P 10	A/P 10	All 10
B5	5	4	4	4	9	7	6	**6**
Buffy	4	4	3	3	7	5	5	**4**
Sim	3	3	3	2	8	8	7	**6**
Sm P	5	4	4	4	9	9	9	**9**
TA	5	4	4	4	9	9	8	**7**
Total	**4.4**	**3.8**	**3.6**	**3.4**	**8.4**	**7.6**	**7**	**6.4**

Table 4: Metric similarity, intersection of top 5 and 10

These results (and results we present below) show no significant difference between the link-based metrics. In-degree and authority are particularly similar. This should be surprising – the primary motivation for the authority algorithm was that in-degree isn't enough, that all links are not equal. Do our results prove this assumption false? No – but they require further consideration.

By starting with items from Yahoo, we almost guaranteed that items in the neighborhood graph we constructed would be relevant to the topic. In contrast, other evaluations of Kleinberg's algorithm [4,7,8] have begun with much noisier neighborhoods. Typically, they've started with a base set of items returned by a search engine, many of which are of dubious relevance, and then added items that link to or are linked to by items in the base set. This sort of neighborhood is likely to contain many pages that are not relevant to the original query. Kleinberg argued that while some of these irrelevant pages have high in-degree, the pages that point to them are not likely to have high out degree; in other words, they don't form a coherent topic. In such cases, the authority/hub algorithm will assign low scores to some items with high in-degree.

To follow through with this argument, we see that two processes are going on: (1) obtaining a set of relevant items, and (2) rating the quality of the items in this set. As commonly conceived, the authority algorithm helps with both. However, our experiment shows that if one already has a set of relevant items, in-degree alone may be just as good a quality measure. Many manually constructed collections of topically relevant items are available from general purpose or topic-oriented directories.

A further note is that the in-degree metric we're using is *site* in-degree. By aggregating links to the site level, we avoid the problems Bharat & Henzinger identified (links between pages that belong to a common site, and mutually reinforcing relationships between two sites). They showed that solving these problems resulted in significant improvements to the basic authority algorithm. The site in-degree metric accrues the same benefits.

4.3 Can We Predict Human Quality Judgements?

We tested how well the rankings induced by each of the features listed in section 3.2 matched expert quality judgements. We wanted to compute the precision of each ranking; to do this, we needed the set of good – high quality – items for each topic. We defined the good items as those that a majority of experts rated as good (i.e., scored 5, 6, or 7). Table 5 shows the total number of items for each topic, number of good items, and proportion of good items. The proportion of good items serves as a useful baseline; it tell us that, across all topics, if you picked a set of 10 items at random, you'd expect about 3 to be high quality.

Topic	total	# good	Proportion good
Babylon 5	40	19	0.48
Buffy	41	15	0.37
Simpsons	39	10	0.26
Smashing Pumpkins	41	7	0.17
Tori Amos	42	13	0.31
Average			**0.32**

Table 5: Number and proportion of good items

For ease of presentation, we present results for 10 metrics. The same 5 metrics performed best in all analyses, so we include them. We also found that all site-based metrics outperformed their url-based counterparts in all cases (e.g., number of images on the entire site was better than number of images on the root page), so we omitted the url-based versions. None of the text relevance metrics performed well, but we include the best – maximum relevance score – for the sake of comparison.

Using the set of good items, we computed the precision at 5 and at 10 for each metric[2]. Table 6 presents the results, with metrics ordered by average precision at 5. The table shows that the top four or five metrics all perform quite well. For example, the in-degree metric has a precision at 5 of 0.76 – on average, nearly 4 of the first 5 documents it returns would be rated good by the experts. This is more than double the number of good documents you would get by selecting 5 at random from the expert dataset. And recall that most of the items in the expert dataset probably are of pretty good quality, since they were selected by multiple subjects in phase 1 of our experiment. Thus, we speculate that in a larger dataset, the improvement in quality obtained by using these metrics is even greater.

Metric		B5	Buffy	Sim	Sm P	TA	Avg
In degree	*at 5*	0.8	0.8	0.8	0.8	0.6	0.76
	at 10	0.6	0.7	0.6	*N/A*	0.5	0.6
# Pages on site	*at 5*	0.8	1	0.6	0.6	0.6	0.72
	at 10	0.8	0.8	0.5	*N/A*	0.4	0.63
Authority score	*at 5*	0.8	0.6	0.8	0.8	0.6	0.72
	at 10	0.7	0.7	0.5	*N/A*	0.5	0.6
PageRank score	*at 5*	1	0.8	0.6	0.8	0.4	0.72
	at 10	0.7	0.6	0.6	*N/A*	0.4	0.58
# Images	*at 5*	1	0.6	0.6	0.6	0.4	0.64
	at 10	0.8	0.7	0.5	*N/A*	0.5	0.63
Out degree	*at 5*	0.8	0.4	0.4	0.4	0.6	0.52
	at 10	0.5	0.5	0.5	*N/A*	0.5	0.5
# Audio files	*at 5*	0.2	0.4	0.6	0.6	0.8	0.52
	at 10	0.2	0.2	0.5	*N/A*	0.6	0.38
Hub score	*at 5*	0.8	0.2	0.4	0.4	0.6	0.48
	at 10	0.4	0.5	0.4	*N/A*	0.5	0.45
Max Rel Score	*at 5*	0.4	0.6	0.6	0.2	0.4	0.44
	at 10	0.7	0.5	0.5	*N/A*	0.4	0.53
Root Page Size	*at 5*	0.6	0	0.4	0.4	0.2	0.32
	at 10	0.5	0.2	0.3	*N/A*	0.2	0.3

Table 6: Precision at 5 and 10

Since the link-based metrics were highly correlated, it should be no surprise that they have similar precision. However, it is surprising how well a very simple metric performs: in this dataset, simply counting the number of pages on a site gives as good an estimate of quality as any of the link-based computations (and number of images isn't bad, either). We speculate that the number of pages on a site indicates how much effort the author is devoting to the site, and more effort tends to indicate higher quality.

The precision analysis abstracted away from the item scores, which could conceal significant differences. For example, suppose that two metrics have identical precision. In principle, they could return completely different sets of items; further, one metric could returned the best – highest ranked – of the good items, while the second returned the worst of the good items. Thus, we wanted to do another analysis using item scores to check for this possibility.

We experimented with two different item scoring schemes, the average of all expert scores and a majority score – (# of experts rating item as good / # of experts rating the item). The two methods yielded similar results, and for the sake of consistency with previous analysis, we used majority score.

Metric		B5	Buffy	Sim	Sm P	TA	Avg
Majority Score	*at 5*	1	1	1	0.9	1	.96
	at 10	1	0.9	0.7	0.7	0.9	.84
In degree	*at 5*	0.8	0.7	0.7	0.8	0.5	.71
	at 10	0.6	0.7	0.6	0.4	0.6	.57
Authority score	*at 5*	0.8	0.5	0.5	0.5	0.5	.69
	at 10	0.7	0.6	0.5	0.4	0.5	.57
PageRank score	*at 5*	1	0.7	0.5	0.8	0.4	.69
	at 10	0.7	0.6	0.6	0.4	0.4	.53
# Pages on site	*at 5*	0.7	1	0.6	0.6	0.4	.66
	at 10	0.8	0.8	0.5	0.4	0.3	.56
# Images	*at 5*	0.9	0.7	0.6	0.6	0.3	.62
	at 10	0.8	0.7	0.5	0.4	0.5	.56
# Audio files	*at 5*	0.3	0.5	0.4	0.6	0.8	.52
	at 10	0.2	0.3	0.4	0.4	0.6	.39

[2] Since there were only 7 high quality items for Smashing Pumpkins, we could not compute precision at 10 for this topic. Accordingly, the average precision at 10 is for the other four topics.

Out degree	*at 5*	0.7	0.4	0.4	0.4	0.5	.49
	at 10	0.5	0.5	0.4	0.4	0.5	.45
Hub score	*at 5*	0.7	0.3	0.4	0.4	0.5	.47
	at 10	0.4	0.5	0.4	0.4	0.5	.44
Max Rel Score	*at 5*	0.3	0.5	0.6	0.2	0.3	.39
	at 10	0.6	0.4	0.5	0.3	0.4	.43
Root Page Size	*at 5*	0.5	0.1	0.2	0.5	0.3	.31
	at 10	0.4	0.2	0.3	0.3	0.3	.28

Table 7: Majority Score at 5 and 10

Table 7 presents the results. For reference, we present the average scores for the top 5 and 10 items as ranked by the expert majority score itself. This is the ideal – no metric can exceed it.. A score of 1 (e.g., for majority score at 10 for Babylon 5) means that all experts rated all items as good. A score of .8 (e.g., in-degree at 5 for Smashing Pumpkins) means that 80% of experts rated all 5 items as good. The best metric is in-degree. It performs about 74% of the ideal at 5, and 68% at 10.

The same metrics – in-degree, authority, page rank, #pages, and #images – are in the top 5 slots in each of the four analyses (precision/majority score at 5/10), although their order varies a little. We wondered whether there were any significant differences among the metrics, so we applied a t-test to each pair of metrics, for each analysis.

Table 8 presents the results of this analysis for majority score at 5 (results were similar for majority score at 10 and precision at 5 and 10). Metrics are ordered by their average majority score at 5; this score is given in the diagonal cells. All other cells contain the p-values returned by the t-test; a p-value is displayed in bold if it indicates a significant difference at the 0.05 level. All the comparisons for a particular metric are found by reading down a column; for example, the comparisons between in-degree and all other metrics are in the first column.

We highlight a few interesting results. First, there were no significant differences between any of the first five metrics. Second, in-degree was significantly better than the rest of the metrics (i.e., other than the top 5). Authority, PageRank, and number-of-pages were similar, except their advantage over #-audio-files wasn't quite significant at the 0.05 level. Third, all of the top 5 methods are significantly better than text similarity. Perhaps text similarity fares so poorly because we started with a set of relevant documents; in other words, if there were more variance in relevance, maybe higher relevance could indicate higher quality.

5. CONCLUSIONS: SUMMARY AND FUTURE WORK

We have investigated the utility of various computable metrics in estimating the quality of web documents. We showed that topic experts exhibit a high amount of agreement in their quality judgements; however, enough difference of opinion exists to warrant further study. We also showed that three link-based metrics and a simple content metric do a very good job of identifying high quality items.

Our results contained two main surprises – first, that in-degree performed at least as well as the more sophisticated authority and PageRank algorithms, and second, that a simple count of the pages on a site was about as good as any of the link analysis methods.

One important area for future work is to carry out the same type of analysis on a larger scale. There are several ways the scope could be enlarged:

- More domains and topics – by investigating topics that don't concern popular entertainment, and more topics of all kinds; we can see whether expert agreement and the performance of various algorithms are influenced by domain or topic.
- More experts – this enables a better determination of the extent to which experts agree and provides a better target for evaluating algorithms; the most plausible way to get more experts is to do a distributed, web-based experiment.

	In degr	Authority score	Page Rank score	#Pages on site	#Images	#Audio files	Out degree	Hub score	Max Rel Score	Root Page Size
In degree	*0.71*									
Authority score	.63	*0.69*								
PageRank score	.70	.93	*0.69*							
# Pages on site	.43	.63	.69	*0.66*						
# Images	.17	.28	.33	.5	*0.62*					
# Audio files	**.01**	.05	.07	.13	.32	*0.52*				
Out degree	**.01**	**.03**	**.02**	**.04**	.09	.76	*0.49*			
Hub score	**0**	**.01**	**.01**	**.04**	.05	.57	.75	*0.47*		
Max Rel Score	**0**	**0**	**0**	**0**	**0**	.15	.16	.33	*0.39*	
Root Page Size	**0**	**0**	**0**	**0**	**0**	**.01**	**.03**	**.02**	.29	*0.31*

Table 8: Statistical Significance for Majority Score at 5 ($p < .05$)

With a larger dataset, other methods can be employed. For example, we tried using a rule learning system to learn rules to predict when an item would be rated as good (or not) by the experts. However, given the relatively small number of items, the learned rules were too specific. In addition, with more data, differences between certain metrics may become statistically significant.

ACKNOWLEDGMENTS

We thank all the participants in our user studies.

REFERENCES

1. Amento, B. User Interfaces for Topic Management of Web Sites. Ph.D. Thesis, Department of Computer Science, Virginia Tech, forthcoming
2. Amento, B., Hill, W., Terveen, L., Hix, D., and Ju, P. An Empirical Evaluation of User Interfaces for Topic Management of Web Sites, in *Proceedings of CHI'99* (Pittsburgh PA, May 1999), ACM Press, 552-559.
3. Baldonado, M.Q.W., and Winograd, T. An Information-Exploration Interface Supporting the Contextual Evolution of a User's Interests, in *Proceedings of CHI'97* (Atlanta GA, March 1997), ACM Press, 11-18.
4. Bharat, K. and Henzinger, M.R. Improved Algorithms for Topic Distillation in a Hyperlinked Environment. ACM SIGIR Conference on Research and Development in Information Retrieval 1998.
5. Buckley, C. Implementation of the SMART Information Retrieval System, Department of Computer Science, Cornell University, 1985, TR85-686.
6. Card, S.K., Robertson, G.C., and York, W. The WebBook and the Web Forager: An Information Workspace for the World-Wide Web, in *Proceedings of CHI'96* (Vancouver BC, April 1996), ACM Press, 111-117.
7. Chakrabarti, S., Dom, B., Gibson, D., Kleinberg, J., Raghavan, P., Rajagopalan, S. Automatic Resource Compilation by Analyzing Hyperlink Structure and Associated Text. *Computer Networks and ISDN Systems* 30 (1998), 65-74
8. Kleinberg, J.M. Authoritative Sources in a Hyperlinked Environment, in *Proceedings of 1998 ACM-SIAM Symposium on Discrete Algorithms* (San Francisco CA, January 1998), ACM Press.
9. Page L., Brin S., Motwani R., and Winograd T. The PageRank Citation Ranking: Bringing Order to the Web. *Stanford Digital Libraries Working Paper.*
10. Pirolli, P., Pitkow, J., and Rao, R. Silk from a Sow's Ear: Extracting Usable Structures from the Web, in *Proceedings of CHI'96* (Vancouver BC, April 1996), ACM Press, 118-125.
11. Pirolli, P., Schank, P., Hearst, M., and Diehl, Scatter/Gather Browsing Communicates the Topic Structure of a Very Large Text Collection, in *Proceedings of CHI'96* (Vancouver BC, April 1996), ACM Press, 213-220.
12. Pitkow, J., and Pirolli, P. Life, Death, and Lawfulness on the Electronic Frontier, in *Proceedings of CHI'97* (Atlanta GA, March 1997), ACM Press, 383-390.
13. Terveen, L, Hill, W., and Amento, B. Constructing, Organizing, and Visualizing Collections of Topically Related Web Resources. *ACM Transactions on Computer-Human Interaction*, 6,1 (March 1999), 67-94.