Improving the Evaluation of Web Search Systems

Cathal Gurrin[1] & Alan F. Smeaton

School of Computer Applications, Dublin City University, Ireland
cgurrin@computing.dcu.ie

Abstract. Linkage analysis as an aid to web search has been assumed to be of significant benefit and we know that it is being implemented by many major Search Engines. Why then have few TREC participants been able to scientifically prove the benefits of linkage analysis over the past three years? In this paper we put forward reasons why disappointing results have been found and we identify the linkage density requirements of a dataset to faithfully support experiments into linkage analysis. We also report a series of linkage-based retrieval experiments on a more densely linked dataset culled from the TREC web documents.

1 Introduction

The first generation of web search engines which have contributed to the huge popularity of the WWW were based on directly computing the similarity between a query and the text appearing in a web page and were effectively a direct application of standard document retrieval techniques. While these initial "first generation" web search engines addressed the engineering problems of web spidering and efficient searching for large numbers of both users and documents, they did not innovate much in the approaches taken to document ranking.

In the past few years we have seen most, if not all, web search engines incorporate linkage analysis as part of their retrieval operation. Anecdotally this appears to have improved the precision of retrieval yet, up until recently, there has been little scientific evidence in support of the claims for better quality retrieval, especially using the conventional TREC evaluation methodology. Participants in the four most recent TREC conferences (1999 - 2002) have been invited to perform benchmarking of information retrieval systems on web data and have had the option of using linkage information as part of their retrieval strategies. Up until TREC-2002, the general consensus was that except in extremely rare cases, and for insignificant improvements anyway, linkage information had not yet been successfully incorporated into conventional retrieval strategies when evaluated using the TREC test collection methodology. In most cases, linkage-based information was found to significantly harm conventional retrieval though improvements had been found specifically in a homepage finding task, which was to locate homepages contained within the two TREC test collections used in 2001 and 2002.

[1] The work presented in this paper is based on research undertaken by the first author as a postgraduate student while working on his Ph.D. dissertation.

In this paper we present a rationale as to why we believe TREC Web Track participants (including ourselves) using the TREC web-based test collections (prior to 2002) have been unable to demonstrate improved retrieval performance when incorporating linkage analysis into retrieval strategies. We begin with a brief introduction to linkage analysis in the next section and follow that with an overview of the TREC web track focusing on an analysis of the datasets employed. We follow this with a description of a densely linked subset of a TREC web collection and our experiments on this subset. Finally we will compare the characteristics of TREC datasets to our own survey of web structure and suggest the linkage requirements for any future datasets to support TREC style evaluation of linkage-based retrieval.

2 An Introduction to Linkage Analysis and its use in Retrieval

The “first generation” of search engine primarily utilised the content of the document when generating ranked listings. However, an additional source of latent information available to web collections is how documents are linked together and it is the study of this aspect of the web that is referred to as linkage analysis. More recent search engines utilise linkage data and are able to gather information mined from the documents themselves, as well as information from the linkage structure of the web. In most cases this linkage information is represented as a ‘Connectivity’ or a ‘Linkage’ Score for each document, which will influence final document ranking.

Generally speaking, on the WWW we can separate links into one of two broad types based on their intended function when created:

- *On-site* links are created to link documents within a particular domain and exist to aid the user in navigating within a domain, or web site. These links are not generally believed to carry much exploitable weight of human judgement.
- *Off-site* (content, or outward) links on the other hand link documents from different domains (across web site boundaries). They are found to mostly link from a source document to a target document that contains similar and, in the web page author's opinion, useful information. The requirement of on-site links to support structural navigation does not apply to off-site links.

In general we can assume that a document with a higher number of off-site in-links (where in-links are hypertext links pointing to a web page) will be a more ‘popular’ document than one with less off-site in-links. For the purpose of linkage analysis we are interested primarily in the number of (and the quality of) off-site citations (in-links) that a web page receives as opposed to on-site citations. If a web page receives a lot of off-site citations then we can broadly conclude that this page may be a better page than one that receives significantly fewer off-site citations. Thus in the context of linkage information for conventional web searching we should primarily be interested in off-site links as opposed to on-site links.

Given that an off-site link to a document can be seen as an indication of the usefulness of that document, a simple linkage score can be generated for each document based on the off-site indegree of that document and hence we can rank documents by off-site indegrees. Researchers at AT&T [1] have demonstrated, using their own crawled data, that incorporating indegree ranking into retrieval strategies is equally as good as incorporating other more advanced techniques for using linkage information, such as the PageRank algorithm that we will now discuss.

The best known linkage analysis technique in use on the web today is probably the PageRank algorithm [2], believed to be implemented in the Google search engine [3]. PageRank is based on a simple indexing-time process that generates a linkage score (the PageRank) for each document in the search engine's index. This PageRank score is combined at query time with other sources of evidence such as a content-only score giving a final document score used in ranking results. PageRank is based on a simulation of a random user's behaviour while browsing the web where the user keeps clicking on successive links at random. Due to the fact that a user can get caught in page loops, rather than looping forever, the user jumps to a random web page (chosen using the vector E over all web pages). E is normally uniform for all web pages.

PageRank is calculated over a number of iterations until an acceptable level of convergence of the PageRanks has been reached. The PageRank (*Pr'*) of a document is calculated using a formula similar to the following (see formula 1), where S_n is the set of documents that link into document *n*, *c* is a constant used for normalisation, Pr_n is the current PageRank of *n* and *E* is a uniform vector over all web pages.

$$\mathrm{Pr}'_n = c \cdot \sum_{m \in S_n} \frac{\mathrm{Pr}_m}{outdegree_m} + (1-c) \cdot E_n \ . \tag{1}$$

Another well-known technique incorporating linkage information in web searching is Kleinberg's [4], which is similar to PageRank in that it is an iterative algorithm based purely on the linkage between documents but it has major differences, namely:

- It is executed at query time, rather than at indexing time
- It computes two scores per document as opposed to one single score. Hub scores reflect a document's usefulness as a source of links to relevant content while Authority scores represent a document's usefulness as a source of relevant content
- It is processed on a small subset of 'relevant' documents, not all documents.

The Hub and Authority scores are calculated for each document on a small subset chosen due to their rank in a content-only search run, or due to their being in the immediate neighbourhood of these highly ranked documents. The process is iterative and the Authority and Hub vectors will eventually converge, at which point the iterations can stop. Once convergence has been achieved, the documents are ranked into two groups, by hub (links to content) and authority (content) scores. A number of improvements to this model have been suggested and successfully evaluated [5].

3 Evaluation of Web-Based Retrieval: The TREC Web Track

From 1999 to 2001 a web "track" in the annual TREC benchmarking exercise has supported participants in their endeavours to find out whether the best methods in ad-hoc (conventional) retrieval also work best on the TREC collections of web data and whether link information in web data can be used to obtain more effective retrieval than using page content alone [6]. In 2002, this ad-hoc search task has been replaced by a Topic Distillation task (evaluated using the measure of precision at 10), the goal of which is to find a small number of key resources on a topic as opposed to the more conventional (ad-hoc) listing of relevant pages. Topic Distillation, although not that far removed from ad-hoc is perhaps more suited to web search evaluation.

In order to support these experiments TREC distributes test collections. TREC test collections consist of three components, a set of documents, a set of queries (called topics) and a set of relevance judgements for each query for use in evaluating the performance of retrieval. The TREC relevance judgements are incomplete (obtained using the pooling method) and are almost always binary relevance judgements.

The first TREC test collection for the web track was the WT2g collection that was used in the TREC-8 conference in 1999. This was a 2GB collection and was said to contain an "interesting quantity" of closed hyperlinks (having both source and target within the dataset). A larger collection of 10 GB of web data, known as WT10g, was used in TREC-9 and TREC-2001. Most recently an 18 GB collection (.GOV) was used for TREC 2002. We examine each collection below.

3.1 TREC-8 (WT2g)

As stated above, the TREC-8 web track used a 2GB collection called WT2g that consisted of 247,491 HTML documents. The WT2g collection is a subset of the 100GB VLC dataset, which is itself a subset of a 300 GB Internet Archive crawl completed in early 1997. Seventeen participating groups took part in the web track in TREC-8 and those that utilised the link information implemented a variety of approaches including Kleinberg's and PageRank mentioned earlier. We ourselves implemented techniques based on citation counting [7].

To the initial surprise of many participants, no participating group (save one with insignificant improvements) managed to improve precision over that attained by conventional content-only searching when linkage information was incorporated into the retrieval process. There were a number of reasons put forward to explain this [8] but the primary reason seemed to be the sparcity of linkage data within WT2g. The number of closed off-site links within WT2g is 2,797 out of 1,166,702, or 0.24% of the total [6], which we, and others, found to be insufficient to support effective linkage-based web retrieval.

3.2 TREC-9 & TREC-2001 (WT10g)

The shortcomings of WT2g led to the creation of a new collection, WT10g, which was used in the web tracks of TREC-9 and in TREC-2001. A corpus size of 10GB was chosen which comprised 1,692,096 documents. Similar to the preceding WT2g, WT10g was also subset of the 100GB VLC dataset but was extracted from the VLC in such a way that maximised the number of off-site links included within WT10g. However, the linkage density of the VLC serves to restrict the number of links that are candidates for inclusion in any extracted collection. The following diagram (Fig. 1) illustrates some of the issues related to the construction of WT10g.

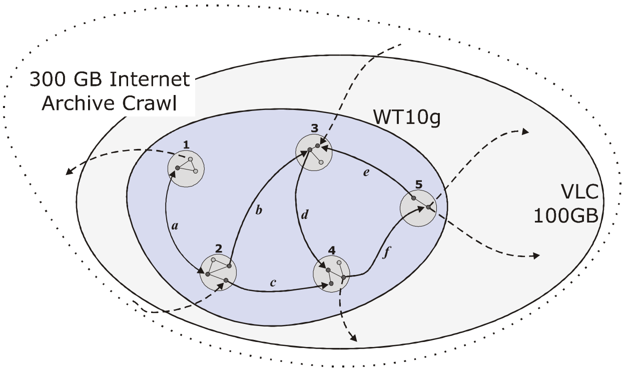

Fig. 1. Construction of the WT10g Collection

Figure 1 shows the original 300 GB crawl, the VLC 100 GB subset of that, then the WT10g subset of the VLC with websites marked 1, 2, 3, 4 and 5, and sample links. The available links in the WT10g collection would be all those contained within websites 1 to 5 (on-site links) as well as those (off-site links) between sites within the WT10g dataset (labeled a to f in Figure 1). The WT10g corpus contained a total of 171,740 off-site links for 1.69 million documents, averaging at 1 off site link for every ten documents. Thus any linkage-based techniques could only influence a small proportion of the documents within WT10g and once again none of the participants' experiments (22 participants in all) for TREC-9 (including our own experiments on citation analysis and spreading activation [9]) were able to illustrate any improvement in performance over content-only retrieval when incorporating any linkage-based algorithm into the retrieval process.

TREC-2001 (the 10th TREC conference) once again encouraged participants (29 in all) to partake in linkage analysis experiments using WT10g and once again linkage analysis was not found to aid retrieval performance in the ad-hoc retrieval task. This time, in addition to the conventional retrieval task, a new task was introduced to the web-track, namely a homepage finding task (essentially a known-item task) which it was hoped would better support linkage-based methods than would conventional IR experiments. Prior evaluation of such methods during construction phase of WT10g had illustrated that such a task could indeed yield performance improvements [10] and some linkage-based improvements were indeed evident in this task.

3.3 TREC-2002 (.GOV)

The .GOV collection used in 2002 consisted of 1,247,753 documents (not all HTML) from a fresh crawl of web pages made in early 2002. It was hoped that fresher web data would better reflect today's WWW. Findings for TREC-2002 illustrate that for some groups the application of linkage analysis did indeed improve retrieval performance in the Topic Distillation task [11] as well as (once again) in the Named Page finding task. So what was the difference between the .GOV collection and the previous collections? The off-site link density of the .GOV collection (averaging 1.98 off-site in-links for each document) was far greater than that of WT10g (0.101). This, we believe, was a primary reason for the encouraging findings.

In summary the web track in TREC has contributed in a major way to focusing attention on scientifically measuring the effectiveness of web searching but progress in actually achieving that measurement has been slow, principally, we believe, because of the test collections used. Our belief is that a dataset better capable of supporting linkage-based web IR would better illustrate the benefits of linkage-based retrieval. The .GOV collection was a step on the way but is not truly representative. The results of experiments on the .GOV collection have only recently become available and the experiments presented in this paper primarily focus on the WT10g (pre .GOV) collection. The recent findings of TREC participants using .GOV do not affect our findings presented herein in any way.

In an effort to generate a more representative collection than the then available WT10g, we generated a new dataset, based on WT10g, but which maximised the density of off-site links and this is described in the following section.

4 Extracting a Densely Linked Subset of WT10g

The creation of the WT10g dataset seriously underestimated the density of off-site links if it was to be reasonably compatible in structure to the WWW and to support faithful experiments on web searching. However, the primary advantage of using WT10g is readily available relevance judgements for 100 queries. We used the TREC-9 query set (50 queries) for our experiments. Generating a new dataset of our own to allow us to do linkage retrieval experiments would require undertaking our own relevance judgements, which is both expensive and time-consuming. Our solution was to develop a subset of WT10g (including relevance judgements) and to evaluate a content-only experiment and to compare all of our linkage experiments against this benchmark content-only experiment, using the relevance judgements.

When generating this densely linked subset of WT10g (called WT_Dense) we had two requirements for the new collection, these being:

1. To maximise the number of off-site links in the dataset, and
2. To maximise the size of the new dataset itself.

Generating a dataset to satisfy these two requirements was straightforward and we simply generated the dataset by following a three-step procedure as described below.

1. We identified all documents that are linked via all 171,740 off-site links (any document linked by a…f in Figure 1).
2. All such documents were extracted to produce a set of 120,494 unique documents.
3. All links (both on-site and off-site) between these documents and the relevance judgements for these documents were extracted from the WT10g collection. It should be noted that we are focusing on off-site links as opposed to on-site links, which we believe to be of lesser importance for linkage-based retrieval.

4.1 Comparing WT_Dense to WT10g

The composition of WT_Dense compared with WT10g is summarised below:

Table 1. Comparing WT10g and WT_Dense

	WT10G	**WT_Dense**
Number of Documents	1,692,096	120,494
Number of off-site links	171,740	171,740
Average off-site indegree	0.10	1.43
Number of unique servers represented	11,680	11,611
Average number of docs per server	144	10
Generality	0.15%	0.21%
Number of TREC-9 Queries with relevant documents	50	36
Average number of relevant documents per query	52	7

As can be seen from Table 1, WT_Dense contains a far higher density of off-site links, an average of 1.43 per document while keeping the generality (percentage of documents relevant to any of the topics) of the dataset similar to WT10g. However, results of a survey of web structure presented later in this paper suggest that this overage indegree figure of 1.43 still falls way below the true figures as found on the web.

As expected the number of servers represented was almost identical in WT_Dense as it was in WT10g, this is because of the inclusion of all off-site links and both the source and target of each link. The one drawback of this is that the average number of documents on each server is only 10 as opposed to 144 with WT10g. This is unavoidable as we only have 7.1% of the WT10g dataset represented in WT_Dense, but it means that we are taking the core pages, home pages and top pages from almost all of the 11,680 servers in the WT10g collection. In addition, fourteen of the fifty TREC-9 queries had no relevant documents in WT_Dense and reducing the number of queries to 36, thus reducing the number of performance comparisons. In addition, although unavoidable, the average number of relevant documents per query was reduced to 7, which we note to be well below the norm.

4.2 Experiments on WT_Dense

We ran a number of retrieval experiments (content-only and linkage based) on the new densely linked dataset using the TREC-9 topics (450-499). We used manually generated queries, which we found to produce the best content-only retrieval performance for the TREC-9 topics.

Our first experiment was a content-only experiment for which we utilised the popular BM25 ranking algorithm. The top 1,000 ranked documents for each query were used as a benchmark against which to compare the retrieval performance of our subsequent linkage-based runs. These subsequent linkage-based runs were based on re-ranking the top 2,000 documents produced in the content-only phase using algorithms based on citation ranking, spreading activation and PageRank.

Citation Ranking

This experiment was based on re-ranking all 2,000 documents from the content-only experiment using both off-site indegrees and the original content-only scores. Before combining both content and linkage scores we normalised the linkage scores so that they would be in an equivalent range to the content-only scores. Let *norm* refer to a normalised score, $indegree_n$ be the off-site indegree of n, Sc_n be the content-only score of n and α be a constant (value of 0.25) used to regulate the influence of linkage evidence (based on values used by AT&T in TREC-9 [12]), we rank by Sc'_n for each document, as shown in equation 2:

$$Sc'_n = Sc_n + \left(norm\left(indegree_n\right)\times\alpha\right)\ . \tag{2}$$

In the results section, this experiment will be referred to as CitationA.

It is our belief that the method chosen for combining linkage and content evidence is essential to successfully incorporating linkage evidence into any retrieval process. We have developed a technique for combining content and linkage scores, which we will refer to as the *Scarcity-Abundance technique*. Essentially this technique dynamically estimates linkage and content influence based on a broadness measure for the topic represented by a user query. The method we employ to do this examines the number of documents retrieved (result set) for each query from a content-only run. A larger result set (the abundance problem) indicates a broader query that would benefit more from linkage-based methods [4], than would a query with narrow focus and vice versa. The content-influence (*1-a*) employed is inversely proportional to the level of linkage influence (*a*). This experiment is referred to in our results as CitationB.

Spreading Activation

Spreading Activation refers to a technique that propagates numerical values (or activation levels) among the connected nodes of a graph. In the context of this experiment it facilitates a document transferring its content-only score across its outlinks. Only documents that are scored in the content-only phase have any effect on the process. The formula for calculating each document score is shown as equation 3. Let S be the set highly scored documents and S_n be the in-set of n, we rank by Sc'_n:

$$Sc'_n = Sc_n + \sum_{m\in S_n} \frac{Sc_m}{outdegree_m}\ . \tag{3}$$

This experiment we shall refer to in our results as SpreadAct.

PageRank Experiment

We also evaluated the PageRank algorithm (an indexing time process) as outlined earlier in this paper. We combined the linkage and content scores for our PageRank experiment using both the parameter technique (linkage weight of 0.25) as well as the scarcity-abundance technique. We will refer to this in our results as PageRankA for the parameter combination and PageRankB for the scarcity-abundance combination.

4.3 Experimental Results

Our linkage experiments are based on re-ranking content-only results, giving us the ability to directly compare linkage and content-only results. Outlined below in a brief summary of our results we see that some linkage experiments actually achieved small improvements in precision (shown as bold/italic in Table 2) over content-only runs when examined at rank positions 5, 10 and 15. This is encouraging because up until these experiments, TREC participants (except one experiment in 1999) were unable to obtain any improvement in retrieval performance when using WT10g data, and although WT_Dense is not the same dataset as WT10g we are using a subset of both the dataset and the relevance judgements and when we ran similar experiments on the full WT10g dataset we did not achieve any improvement in performance at all.

Table 2. Precision Values for the Experiments on WT_Dense

	CONTENT	CITATION A	CITATION B	SPREADACT	PAGERANK A	PAGERANK B
5	0.2389	***0.2444***	***0.2500***	0.0500	***0.2444***	***0.2444***
10	0.1833	0.1833	***0.1861***	0.0639	0.1833	0.1806
15	0.1611	***0.1630***	***0.1630***	0.0722	***0.1630***	0.1593
20	0.1500	0.1472	0.1486	0.0750	0.1486	0.1486
30	0.1167	0.1148	0.1148	0.0787	0.1130	0.1139
100	0.0444	0.0444	0.0442	0.0406	0.0444	0.0444
200	0.0246	0.0247	0.0249	0.0244	0.0246	0.0246
500	0.0114	0.0114	0.0114	0.0114	0.0114	0.0114
1000	0.0061	0.0061	0.0061	0.0061	0.0061	0.0061

Our experiments suggest that by increasing the density of off-site links within a collection one can support successful linkage-based retrieval experiments. Although the improvements are quite small and on a small collection, our findings are backed up by the recent successful results of some TREC participants using the more densely linked .GOV collection in the Topic Distillation task.

Our next step was to examine the linkage structure of the web and compare it to the TREC collections and our own WT_Dense collection. If the linkage density on the web is underestimated within these collections, then it is likely that a more representative dataset would illustrate more clearly if and by how much, linkage-based techniques can aid conventional retrieval performance. In order to evaluate this and to identify the requirements for a small-scale test collection, which faithfully models the linkage structure of the web, we conducted a survey of 5,000 web pages.

5 Examining Real-World Web Structure

We present the results of a small survey of web pages from early 2002 whose aim was to explore the linkage structure of the WWW. In order to estimate in-degree densities we had to examine all out-links from the web pages and since the web is a directed graph and each out-link is also an in-link, observing the average outdegree of each document allowed us to identify the average indegree of every document. This assumes that one does not to include any broken links in this calculation.

However, we need to know more than the average outdegree of each web page as we also need to know the average number of off-site out-links and the average number of on-site out-links which can be discovered by observation.

5.1 A Survey of Web Pages

In order to correctly identify the average in-degree (both off-site and on-site) of web pages we carried out our own survey of the linkage structure of the WWW as it is in 2002 by sampling web pages at random using a random URL generation tool.

Previous Work

Previous work has been carried out in this area, an example being the SOWS III survey of web structure, which also involved the random sampling of WWW pages [13]. The SOWS III survey was based on fetching and examining 200 documents and the findings indicate that the average number of links parsed from each document was 22.9, of which 1.3 were found to be dead, giving an average of 21.6 valid out-links from every document.

In addition to SOWS III, Cyveillance Inc. carried out experiments in 2000 [14] in order to size the Internet and they found that the average page contains 23 on-site out-links and 5.6 off-site out-links, however they did limit their processing of web pages to pages under 200KB in size.

From our point of view, we were interested examining the valid (non-broken) off-site and on-site link structure of documents in order to estimate the number of valid links that exist on the real web so we conducted our own survey.

Surveying the WWW

When generating our own sample we identified the target population as being all web pages that are reachable by a conventional web crawler. Therefore we feel that it is acceptable to base a sample on documents chosen at random from a large crawl of web data, for example, a search engine's index as opposed to truly random URL sampling based on selecting random IP addresses. From [15] we know that 8.24% of web pages are not connected into the main body of the web and if pages from these sections are not manually submitted for inclusion in a crawl, many will not be found and therefore it is acceptable that these pages would not form part of the target population. Hence, the use of crawler based sampling is reasonable and acceptable for the purposes of our experiment.

Our approach to generating random URLs required the use of a web accessible random URL generator [16]. Our sample size was 5,000 and if we examine our sample at a 95% confidence level, our confidence interval is 1.39%, which compares favourably with a confidence interval of 6.93% for SOWS III. This means that we can be 95% confident that our results are +/- 1.39% of our stated figures. One caveat with our survey is that these confidence figures assume truly random sampling and since we relied on URouLette for our random URLs, we are not sure how random our sample is and from how large a list of candidate URLs the random URLs are chosen. All sampling techniques that rely on choosing a document at random from a web crawl are only random with respect to the crawled data and thus cannot be classified as truly random. Truly random sampling would, as mentioned, likely involve some form of random IP address generation and subsequent web page selection.

Observations from our Survey

Based on our survey we are in a position to present our findings on the link structure of these 5,000 documents. We can see from Table 3 that 3,940 documents contain out-links, with only 2,706 containing off-site out-links.

Table 3. Basic linkage structure of documents from the random sample

Documents that contain out-links	3,940
Documents with no out-links	1,060
Documents with off-site out-links	2,706
Documents with on-site out-links	3,571

We estimate the average (HTTP links only) outdegree of WWW documents to be 19.8, which is comprised of 5.2 off-site links and 14.6 on-site links. However, after downloading and examining all of the target pages (to identify broken links) from all links found in the 5,000 document sample we found that 3.2% of all out-links were broken links in that they yield only an HTTP 404 type error or equivalent.

Considering this information we can identify the following valid (non-broken) out-link structure for all documents on the web (rounded to one decimal place) as:

Table 4. Average document outdegree figures from our survey

Average off-site out-degree for each document	4.9
Average on-site out-degree for each document	14.2
Average out-degree for each document	19.1

If we compare our findings to those of SOWS III and Cyveillance we can see that our average outdegree figure is not that dissimilar to either, with SOWS III estimating the average outdegree to be 21.6. However Cyveillance is somewhat higher at 28.6, but we must remember that Cyveillance restricts the size of documents processed and we ourselves only process valid HTTP links. Given that our findings are very similar to SOWS III (and similar to Cyveillance w.r.t. off-site links more so than on-site links) this adds weight to our findings and our belief that utilising URouLette did not affect the randomness of survey to any notable extent.

However, simply examining the average number of links within the collection is not an ideal measurement as the distribution of the indegrees will not be uniform and gives us "little insight into the topology" [17]. An additional test that we can apply to our sample is based on examining the distributions of document outdegrees. If we conclude that the distribution of outdegrees approximates a power-law distribution then this adds more weight to the accuracy of our survey.

5.2 Examining the Distribution of Web Page Outdegrees

It has been discovered that the distribution of web page indegrees follows closely to a power-law distribution [15], yet the distribution does not follow a pure power-law [18], and rather we consider it to approximate a power-law distribution. We are told that the “distribution of inbound links on the web as a whole is closest to a pure power-law” while “category specific distributions exhibit very large derivations from power-law scaling” [18]. This raises issues for the generation of test collections because any attempt to influence the documents comprising a dataset in order to include some category specificity (perhaps to aid in query selection) will result in problems when trying to recreate the natural web link structure. However, test collections using the TREC methodology are non-category specific and thus we can be satisfied that the distributions of document indegrees should approximate a power-law distribution, which is indeed the case for WT10g and .GOV[17].

Power Law Distributions

Power-laws are used in mathematics when one wishes to relate one quantity to the power of another. A power-law implies that small occurrences are extremely common whereas large occurrences are extremely rare, so if applied to web page indegrees or outdegrees this means that the vast majority of web pages have a very small number of in (or out) -links and a few pages have a large number of in (or out) -links.

Power-law distributions are not just used to describe the indegrees of web pages (or computer science problems in general), rather they are common to both man made and naturally occurring phenomena [19]. From computer science we see power-law distributions in web page indegrees [15], outdegrees [20], in the number of pages on websites [21], in Internet growth models [19], and in the distributions of word frequencies in language [22].

The characteristic signature of data that follows a power-law is that when the data is plotted on a log-log scale, the distribution shows itself to be linear (with the slope based on the exponent). Were our sample of web pages to be accurate then the distributions of our outdegree calculations would have that same characteristic signature. In addition, we can calculate the correlation co-efficient, which will range from 0..1, with 1 being perfect positive correlation to a power-law distribution.

If we examine the distribution of non-broken off-site out-links in our web sample then we can see that this approximates a power-law as can be seen in Fig. 2.

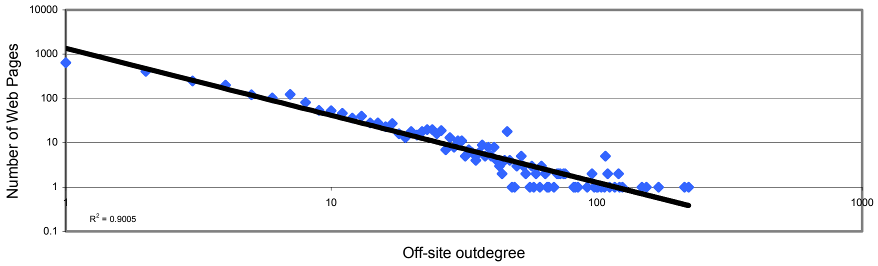

Fig. 2. The off-site outdegree distribution of our sample plotted on a log-log scale with broken links removed, including trendline (correlation co-efficient = 0.9005)

If we examine the distribution of on-site outdegrees of non-broken on-site links we find a power law distribution also, although the correlation co-efficient at 0.8542 is less than that of the off-site distribution.

By examining the distribution of outdegrees from our sample, we have illustrated that both distributions approximate a power-law and that this is precisely what we would expect to find if our sample accurately reflected the web's structure.

We know [15] that document indegrees (including off-site indegrees) also approximate (or follow) a power-law distribution. Consequently, when building a dataset to support faithful experiments into linkage-based retrieval of web pages, the indegree distribution of links within the dataset should approximate a power-law distribution and based on our average outdegree figures from our web page sample we can identify the link density that we would expect to find in such a dataset.

6 Conclusions and Future Work

In this paper we have presented the results of experiments into linkage-based web IR using both the TREC test collections and our own densely linked subset of a TREC test collection. As a result of our disappointing findings and those of other TREC participants we examined the structure of the real web in order to identify the required linkage properties of such a representative test collection, which we present below.

Our belief is that only a collection that accurately recreates the linkage structure of the web will be able to truly answer the question of whether or not incorporating linkage evidence aids retrieval performance for web search and since we do not have such a collection more evidence is needed. We now illustrate the collection requirements.

6.1 Requirements for a Representative Test Collection

As a result of our survey of 5,000 web pages and the work of the TREC web track organisers into methods of constructing a test collection [10] we can identify the requirements for a test collection to support truly accurate investigation into linkage-based retrieval. This test collection should model real web searching, by means of:

- A sufficiently large and representative document set [10].
- A representative link structure within this document set.
- A large set of representative web queries [10].
- Sufficiently complete relevance judgments [10].
- Sufficiently high generality of the dataset so as to clearly illustrate any benefit which linkage-based retrieval techniques bring to web retrieval.

As stated above, the ideal web test collection should include a link structure that accurately reflects the true link structure of the WWW to enable meaningful experiments into linkage-based IR. This link structure can be summarised thus:

- Must have an average off-site indegree of (or adequately near) 4.9.
- Must have an average on-site indegree of (or adequately near) 14.2.
- The indegree distributions (both off-site and on-site) must approximate a power-law distribution with exponents capable of producing appropriate indegree figures

Our findings illustrate that our subset of WT10g, WT_Dense, like the other collections used for web search experiments (even .GOV), seriously underestimate the off-site link density of the WWW. WT10g and WT2g have an even lower off-site link density than WT_Dense or .GOV so the problems with using WT10g and WT2g are even more acute.

Assuming that all out-links (with the exception of broken-links) also act as in-links then Table 5 compares the average off-site indegree figures for the TREC collections, WT_Dense and what our survey suggests to be the ideal figure.

Table 5. Comparing our sample to recent TREC collections

	WT2g	**WT10g**	**WT_Dense**	**.Gov**	**Web Survey**
Average indegree	0.011	0.101	1.425	1.98	4.916

This analysis leads us to conclude that previous and present TREC test collections have seriously underestimated the required link density of off-site links, even though the distribution of these links do follow a power law distribution [17]. Although the TREC web track has aided the research field immensely by providing a framework for experimentation, unfortunately the collections used have not supported TREC participants' experiments into linkage-based web IR sufficiently.

The recent .GOV collection is a major improvement, but still falls short of the required off-site link density while WT_Dense also underestimates the off-site link density by a factor of over three. This leads us to question if there is a certain (critical) density of off-site links required within a collection before it can support linkage experiments. Results from this year's web track suggest that an average of 1.98 off-site in-links into each document is sufficient to show improvements in retrieval performance and our experiments suggest that 1.425 may also be sufficient, but both fall short of the ideal. The benefits (or otherwise) of incorporating linkage analysis into web search will be more clearly illustrated on a more representative collection.

6.2 Future Work

One issue that we have not addressed in this paper is that of how to generate such a test collection. This is a complex issue and one that we plan to examine in the near future. Simply crawling a set of web documents is dependent on the queuing algorithm in the crawler and the results will vary greatly making it an extremely difficult task, hence our current thinking is that the dataset should be generated by carefully selecting documents from a larger collection in order to construct a small-scale representative collection, similar to our generation of WT_Dense. We refer the reader to the paper describing the construction of WT10g [10] for an example of one such process, albeit with differing requirements to those we have presented above. The size of this super-set of documents is subject to experimentation, as it must contain a high enough density of links between documents to enable successful subset extraction. Other issues such as the effects on a test collection by judiciously adding documents to meet a required link distribution and density have yet to be seen. It is with the aim of solving these problems that we continue our research.

References

1. Amento, B., Terveen, L. and Hill, W.: Does 'Authority' mean quality? Predicting Expert Quality Ratings of Web Document. Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in IR (2000)
2. Page L., Brin S., Motwani R. and Winograd T.: The PageRank Citation Ranking: Bringing Order to the Web. Stanford Digital Libraries working paper (1997) 0072.
3. Brin S. and Page L.: The Anatomy of a Large-Scale Hypertextual Web Search Engine. Proceedings of the 7th International WWW Conference (1998)
4. Kleinberg, J.: Authorative Sources in a Hyperlinked Environment. Proceedings of the 9th ACM-SIAM Symposium on Discrete Algorithms (1998)
5. Bharat K. and Henzinger M.: Improved Algorithms for Topic Distillation in a Hyperlinked Environment. Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in IR (1998)
6. Hawking D., Voorhees E., Craswell N. and Bailey P.: Overview of the TREC-8 Web Track. Proceedings of the 8th Annual TREC Conference" (1999)
7. Gurrin, C. and Smeaton, A.F.: Connectivity Analysis Approaches to Increasing Precision in Retrieval from Hyperlinked Documents. Proceedings of the 8th Annual TREC Conference", November 16-19 (1999)
8. Hawking D.: - Overview of the TREC-9 Web Track. Proceedings of the 9th Annual TREC Conference", November 16-19 (2000)
9. Gurrin, C. and Smeaton, A.F.: Dublin City University Experiments in Connectivity Analysis for TREC-9. Proceedings of the 9th Annual TREC Conference" (2000)
10. Bailey, P., Craswell, N. and Hawking, D.: Engineering a multi-purpose test collection for Web retrieval experiments. Information Processing and Management (2001)
11. Wu, L., Huang, X., Niu, J., Xia, Y., Feng, Z., Zhou, Y.: FDU at TREC 2002: Filtering, Q&A, Web and Video Tasks. Draft Proceedings of the 11th Annual TREC Conference, November 19-22 (2002)
12. Singhal, A. and Kaszkiel, M.: AT&T at TREC-9. Proceedings of the 9th Annual TREC Conference, November 16-19 (2000)
13. SOWS III: The Third State of the Web Survey, Available online at URL: http://www.pantos.org/atw/35654-a.html. (last visited November 2002).
14. Murray B. and Moore A.: Sizing the Internet - A White Paper. Cyveillance, Inc., 2000. Available online at URL: http://www.cyveillance.com/web/corporate/white_papers.htm. (last visited November 2002)
15. Broder A., Kumar R., Maghoul, F., Raghavan P., Rajagopalan S., Stata R., Tomkins A. and Weiner J.: Graph Structure in the Web. Proceedings of WWW9 (2000)
16. URouLette Random Web Page Generator, Available online at URL: http://www.uroulette.com. (last visited November 2002)
17. Soboroff, I.: Does WT10g look like the Web?. Proceedings of the 27rd Annual International ACM SIGIR Conference on Research and Development in IR (2002)
18. Pennock, D., Flake, G., Lawrence, S., Glover, E. and Giles, C.: Winners don't take all: Characterising the competition for links on the web. Proceedings of the National Academy of Sciences, Volume 99, Issue 8, (April 2002) 5207-5211
19. Mitzenmacher M.: A Brief History of Generative Models for Power Law and Lognormal Distributions. Allerton (2001)
20. Faloutsos, M., Faloutsos, P., Faloutsos, C.: On Power-Law Relationships of the Internet Topology. Proceedings of ACM SIGCOMM 99 (1999)
21. Adamic, L. and Humberman B.: The Web's Hidden Order. Communications of the ACM, Vol. 44, No. 9 (2001)
22. Adamic, L.: Zipf, Power-laws, and Pareto - a ranking tutorial. Available online at URL: http://www.hpl.hp.com/shl/papers/ranking/. (last visited November 2002).