

Link analysis and Pagerank

Sjors E.F. van Berkel
Student number: 1262882

Abstract

Indexing the Internet, and automatically recognizing the importance of pages. It is one of the main issues of Information Retrieval about the Internet. In this paper we will look at two link analysis techniques that try to do exactly this by looking at the hyperlink structures between pages. We will see what the idea behind these techniques is, and why they work, or why they sometimes fail to work. We continue by looking at advanced techniques that leverage the underlying implementations to tweak the search results, to make them even more specific.

1 Introduction

The main issue in Information Retrieval is the following: given a query and a system with documents, return the documents in the system that are most relevant to the query. The Internet is a relatively young field, and its size and structure are unprecedented. Because of this, it poses a lot of new challenges for the field of Information Retrieval.

Because of the vast majority of web pages on the Internet, the search paradigm is possibly the single most important factor for content discovery on the Internet. Successfully providing the user with content he or she searches for is however, not trivial. The strength and weakness of the Internet is that everyone can author web pages and publish them at virtually zero cost. This creates the possibility for anyone to publish information, generating a combined information source unprecedented in size or diversity. Combine this with the hypertext protocol, which only includes a minimal set of semantics for content description, and we can see a great challenge in finding relevant information on the Internet.

An interesting feature of Internet pages are the hyperlinks that create a unidirectional link from one page to the other. This feature effectively means we can render web pages as vertices in a graph, connected by the hyperlinks that function as directed edges. The downside of these links is that they only carry a small text describing the link (in the most positive case). They have no real meaning when considered separately, they only create a sense of the most general relation between two web pages.

In this paper, we look at link analysis techniques that try to infer the importance of pages, based on their relationships, and use this information to determine the relevance of web pages to a query. We will look at the a paper by J.M. Kleinberg[Kle99], describing a method for link analysis, and the paper on PageRank analysis[PBMW98], an important component of the Google search engine.

2 Link Analysis

The main idea of both [Kle99] and [PBMW98] is to deduce a sense of authority of web pages on a subject by looking at what other pages have linked to them (so-called backlinks of a page). We will now look at some specific aspects of the two systems to get an idea of how they work.

2.1 PageRank

PageRank[PBMW98] maintains a score of each page by adding up the scores of the backlinks of that page. The score of a backlink is the score of the page that contains the link, divided by the number of links that

page has. It is easy to see that the definition is recursive, and hence it takes a number of iterations let the scores converge to a stable state. A powerful aspect of the PageRank method is that it works without the need to make any assumptions, because pages that are linked to very often get authority, and they also automatically transfer authority to the pages they link to.

To test PageRank in [PBMW98], the researchers implemented a simple title based search engine called Google, and ordered the results according to their PageRank scores. Even though this idea was simple, it already worked remarkably well (it returned a lot of prominent results first) for a search engine in 1998.

2.2 Kleinberg

The link analysis algorithm described by [Kle99] works slightly differently. The algorithm tries to separate internet pages that function as hubs and pages that function as authorities. Hubs are human moderated portal pages that contain links to a large number of different subjects. The hubs link to pages that, for a certain subject, are likely to be (semi-)authoritative pages on that subject. Hub pages can be recognized because they have a lot of outgoing links to authoritative pages, and they are generally not linked to by authoritative pages. Authoritative pages on the other hand, have a lot of incoming links from hubs.

Instead of assigning a score to each page, like PageRank (Section 2.1) does, this algorithm assigns an *authority weight* (or x), and a *hub weight* (or y) to every page. If a page p has a lot of links to pages with a high x value, it should have a large y value, and if p is pointed to by many pages with large y -values, it should have a large x -value. This algorithm is also recursive, and iteratively runs this calculation over all the pages, calculating x and y alternately. After each iteration, x^2 of all pages combined is normalized to 1, to make sure the values do not get infinitely large or small. The same is done for y^2 .

To test the results of the algorithm, the researchers typed in a query on search engine Altavista, providing them with a set of search results S . To increase the chance of relevant results, the set T was created, which contained all results of S , and to an unspecified threshold, all pages that were linked to in S . The returned results were also very good, in comparison to the original ordering of results by Altavista.

2.3 Problems of link analysis

One issue [PBMW98] mentions, are so-called dangling links (Section 2.7): links to a page without outgoing links. Although this situation could exist on the “real internet”, it is mostly an artifact of not having downloaded all pages that need to be evaluated. This issue arises because it is virtually impossible to download all pages on the internet, because of its size. According to the writers you can drop these dangling links when calculating the PageRank scores, and add them back in later on. This slightly affects the PageRank scores for the rest of the system, but according to the writers “this should not have a large effect”.

Another issue that both [Kle99] and [PBMW98] mention, is that link analysis works best on queries that will have a lot of results. For more specific queries, [PBMW98] (Section 5.2) proposes merging the ranks as calculated by PageRank with ranks calculated by traditional information retrieval scoring methods. However, they mention it is a “very difficult problem”, and needs “considerable additional effort” in their Google system (at that time). [Kle99] (Section 6) mentions the effect of “diffusion”: the algorithm finds a set of hubs and authorities that are not authorities on the original topic, but rather on a generalization of the topic. They propose multiple strategies for solving this issue, also mixing lexical analysis and their scoring algorithm (in the same sense as PageRank). For example, they propose to measure term frequency in a set of related results to determine their relevance, and incorporate these scores into the final scores (page 24).

Another problem that was mentioned in [Kle99] were the so-called term mixtures (page 25). This problem arises when a user searches for multiple terms; the default results as provided by the algorithm are unlikely to contain information on multiple terms, they will probably only focus on one of the terms. As we will see in Section 3, the algorithm of [Kle99] is able to decompose the results into multiple sets, and from there, it is possible to upgrade the scores of result sets that have more relevance to multiple terms than other result sets.

3 Zooming in: Going Beyond Default Results

An important ability of the algorithms is that they both have possibilities to influence the result sets. [Kle99] (Section 3) uses decomposition of the (original) results matrix into eigenvalues and their eigenvectors. The pairs of hubs and authorities associated with an eigenvector appear to be closely related, and hence can represent a different community than pages associated with another eigenvector. This can happen when, for example, the search terms have multiple meanings, or multiple fields in which they apply. This possible subdivision into multiple result sets provides the search engine with a large spectrum of extra possibilities. The system can do all sorts of inter-resultset comparisons, subtractions, additions, leading to better specified, or even personalized results. For example, imagine a search engine that has knowledge of the user's preferences; it can calculate which of the result sets is closest to the user's preferences, and show boost this result set in the total results. Obviously, such an example only scratches the surface.

[PBMW98] uses another method for influencing the search results (Section 6). The computation of PageRank could be viewed as a random walk over the graph of the web. During this walk, a user can get bored with the current path, and randomly jump to another page, unrelated to his current path. The pages a user can jump to are maintained in a vector. Normally this vector is unbiased; the user can jump everywhere equally with the same chance. If we adjust this vector, and thus adjust the paths a user is likely to walk, we adjust the computed PageRank scores. This provides possibilities for personalization of search results, because we can model the PageRank algorithm to be more likely to visit the pages the user visits, providing results that are closer to the user's preferences. Personalization is again only an example, because the method provides the means for a lot of different adjustments of the PageRank algorithm. One problem (that I identified myself) with this approach, in which it vastly differs from the Kleinberg algorithm, is that PageRank scores will have to be recalculated for every change in this vector, which can take up a lot of time and memory.

4 Conclusion

Having looked at the two link analysis techniques, we can see that they definitely provide a good tool for a lot of search problems. Their power is based upon one of the (few) keystones of the Internet, and at the same time they are so simple to understand, making them powerful in applicability, but not at the expense of simplicity.

Both techniques, at the time of writing, had some issues to iron out, but for the general case, they worked very well. Also, the fact that both papers gave attention to customization of search results is very promising because I think it is important for these results to be useful in different fields.

It would be interesting to see if the current Internet landscape (size and structure), still supports these simple indexing and calculation steps, or that it has just become too large to do this with normal machines.

For research ends, it is too bad the Google search engine is not open source, because it would be really interesting to take a look at the developments that have taken place between the time of writing and now. For example, it would be nice to see if the PageRank algorithm has significantly changed since then.

References

- [Kle99] J.M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*, 46(5):604–632, 1999.
- [PBMW98] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank Citation Ranking: Bringing Order to the Web. 1998.