Age Dependent Document Priors in Link Structure Analysis

Claudia Hauff[1] and Leif Azzopardi[2]

[1] University of Magdeburg, Magdeburg, Germany
hauff@student.uni-magdeburg.de
[2] University of Glasgow, Glasgow, United Kingdom
leif@dcs.gla.ac.uk

Abstract. Much research has been performed investigating how links between web pages can be exploited in an Information Retrieval setting [1, 4]. In this poster, we investigate the application of the Barabási-Albert model to link structure analysis on a collection of web documents within the language modeling framework. Our model utilizes the web structure as described by a Scale Free Network and derives a document prior based on a web document's age and linkage. Preliminary experiments indicate the utility of our approach over other current link structure algorithms and warrants further research.

1 Introduction

Recently, Scale Free Networks (SFN) have been proposed to account for the evolving nature of many real networks based on two factors; the growth of the network and preferential attachment [2]. Such networks are characterized by a power law distribution. It was shown that the World Wide Web is a SFN where the number of links pointing to (in-links) and from a web page (out-links) follow power law distributions [3]. We attempt to utilize the SFN when estimating a document's importance given its age and link information, by exploiting the property that as a page ages (and more pages enter the network) it would attract more links (preferential attachment). Our approach is unlike the traditional link structure analysis algorithms, such as HITS and PageRank, which view the web as a static structure not evolving over time. Whilst a modified version of the PageRank algorithm, that heuristically boosts younger pages with higher scores has been proposed [1], under the SFN approach page age is accounted for in a principled manner. We present a brief overview of our approach and provide some experimental results.

2 Popularity Within a SFN

Briefly, the web as a SFN: starting at time $i = 0$ with $m_i > 0$ web pages, at each time step $i = j$ a new web page $d_{i=j}$ is introduced to the network with m links pointing to different pages already in the network. The probability that an

D.E. Losada and J.M. Fernández-Luna (Eds.): ECIR 2005, LNCS 3408, pp. 552–554, 2005.

existing page d attracts one of these new links is denoted by $\prod(d)$ and depends on the number of links l_d that d has already acquired, such that:

$$\prod(d) = \frac{l_d}{\sum_{d'} l_{d'}} \tag{1}$$

This allows us to derive a function that determines the number of in-links a web page "should" have collected at any given time step i, given the page's age. The expected number of in-links $e_j(i)$ at time i for a page d_j introduced to the network at time j is given by:

$$e_j(i) = m\sqrt{\frac{i}{j}} - m \tag{2}$$

For a collection of web pages the constant m is the average number of out-links and the order in which the pages enter the network is established by ranking the pages according to their age. This expectation can be exploited in deriving a popularity score based on comparing the actual number of in-links with the expected number of in-links for a particular page. The rationale is that we would anticipate a popular page to have more in-links than expected and vice versa for an unpopular page. We obtained the popularity score based on a smoothed ratio of actual over expected number of in-links and normalized to a range between 1 and 3.

3 Experiments and Results

The Language Modeling framework [4] offers a principled way to incorporate query-independent knowledge in a retrieval model. In this framework, a document d is sampled with the prior probability $p(d)$, then from d the query q is drawn with probability $p(q|d)$. Essentially, the joint probability of d and q is used to rank the documents and the prior $p(d)$ allows us to encode its importance. In this study, we compare several different priors in an ad hoc retrieval web task - the Uniform prior, the Document Length prior, the Laplace-smoothed In-link prior, and the PageRank based prior (see [4, 5] for more details) - against our SFN prior. The SFN prior was created by normalizing the popularity scores. The age of each web page was defined as the difference between the last modified date and the current date. Of course, this is not the true age of the page, but a reasonable estimate given the data available. We performed this pilot study on the WT2g Collection, and used the titles of TREC topics 401-450 for evaluation purposes. To compute the query likelihood $p(q|d)$, we used Bayes Smoothing with a Dirichlet prior fixed at 1000.

We report the mean Average Precision for each of the document priors (Table 1). Besides clear ranking $p(q, d)$, the best performing interpolated retrieval value is also presented, where α is the interpolation ratio: $\alpha p(q|d) + (1 - \alpha)p(d)$. The percentage change was computed relative to the uniform prior.

Table 1. Performance of variable document priors on WT2g

Document Prior	Clear mAP	±%	Interpol. mAP	α	±%
Uniform	29.325	-	-	-	-
Document Length	31.395	+7.01	32.544	0.60	+11.02
In-Link	22.918	-21.82	29.277	0.95	-0.13
PageRank	22.639	-22.77	29.270	0.90	-0.15
SFN	29.625	+1.06	29.718	0.60	+1.37

4 Discussion and Conclusions

From our results it is clear that the performance of link priors based on a static view of the network (In-Link and PageRank) is substantially (clear ranking) or slightly worse (interpolation ranking) than the Uniform prior. However, whilst not statistically significant, the SFN prior shows promise with over one percent increase in mean Average Precision over the Uniform prior and other link priors[1]. We believe this provides an encouraging platform from which to develop the model further. Future research will be aimed at addressing several key issues of the proposed method and the limitation of this study. These include: testing on larger web collections where the link structure is more representative of the true web, different ways to generate popularity scores given the expected number of in-links, improvement in the estimation of document age, and application to web retrieval tasks other than ad-hoc retrieval. [2]

References

1. R. Baeza-Yates, F. Saint-Jean, and C. Castillo, *Web structure, age and page quality*, 2nd International Workshop on Web Dynamics (WebDyn 2002), 2002.
2. A.-L . Barabási, R. Albert, and H. Jeong, *Mean-field theory for scale-free random networks.*, Physica A **272** (1999), no. 173; cond-mat/9907068.
3. A.-L. Barabási, R. Albert, and H. Jeong, *Scale-free characteristics of random networks: the topology of the world wide web*, Physica A **281** (2000), no. 69.
4. W. Kraaij and T. Westerveld, *Tno/ut at trec-9: How different are web documents?*, Proceedings of the ninth Text Retrieval Conference TREC-9, 2001, pp. 665–671.
5. D. R. H. Miller, T. Leek, and R. M. Schwartz, *A hidden markov model information retrieval*, 22nd Annual International ACM SIGIR conference on Research and development in information retrieval (California, US), ACM Press, 1999, pp. 214–221.

[1] For brevity we ignore any discussion about the document length prior. However, a report containing full details about this study is available from the first author's web site (http://www.uni-magdeburg.de/hauff).

[2] This work was performed during a 6 month visitation to the Glasgow IR Group and the authors would like to thank Professor C.J. van Rijsbergen for his support and input.