Mining Neighbors' Topicality to Better Control Authority Flow

Na Dai, Brian D. Davison, and Yaoshuang Wang

Department of Computer Science & Engineering, Lehigh University, USA
{nad207,davison,yaw206}@cse.lehigh.edu

Abstract. Web pages are often recognized by others through contexts. These contexts determine how linked pages influence and interact with each other. When differentiating such interactions, the authority of web pages can be better estimated by controlling the authority flows among pages. In this work, we determine the authority distribution by examining the topicality relationship between associated pages. In addition, we find it is not enough to quantify the influence of authority propagation from only one type of neighbor, such as parent pages in PageRank algorithm, since web pages, like people, are influenced by diverse types of neighbors within the same network. We propose a probabilistic method to model authority flows from different sources of neighbor pages. In this way, we distinguish page authority interaction by incorporating the topical context and the relationship between associated pages. Experiments on the 2003 and 2004 TREC Web Tracks demonstrate that this approach outperforms other competitive topical ranking models and produces a more than 10% improvement over PageRank on the quality of top 10 search results. When increasing the types of incorporated neighbor sources, the performance shows stable improvements.

1 Introduction

People inherit reputations from other members in their social network. For example, we say that a professor is an expert in artificial intelligence (AI), it is probable that her advisees get some prestige in the field of AI from her. However, if one of her advisees is an expert in system and computer network rather than AI, the chance that he inherits authority from his advisor is much less than other advisees who are well-known for AI, since the authorities of advisor and advisees within the same field are more likely to influence each other. In addition, the reputation of a researcher can come from other sources besides the advisor. For instance, collaborating with other famous researchers can improve one's reputation; advising a good student who becomes famous afterwards can also improve one's reputation.

Thus, one may think that one's reputation is aggregated from a comprehensive environment, and the more similar people are the greater the influence on one's reputation. These intuitions are also applicable to the interconnected pages on the web, especially benefiting the estimation of web page authority. However, traditional link analysis algorithms seldom consider these two issues at the same time. Algorithms such as PageRank and HITS assume that the authority flow among pages is equally distributed following limited directions. The authority propagation is independent of the contexts correlating to the pages within limited relationships. Although topical ranking models [2, 5] differentiate the page authority within diverse topical domains/communities, none of them directly compared the topicality between two linked pages in authority propagation.

Nie and Davison [4]'s Heterogeneous Topical Rank incorporates such topical context comparison in fine-grained local authority flows within pages. However, the direction of authority flows is only from source to target pages. Other approaches using relevance propagation use neighboring nodes' relevance score to influence the current node for given queries or perform content diffusion among pages to achieve better representation. Qin et al. [7] proposed a generic relevance propagation framework and did a thorough comparison study in this field. Shakery and Zhai [9] proposed a probabilistic framework which propagates page relevance scores for a given query through page inlink and outlinks. However, the relevance propagation sometimes has to proceed online, potentially making the computation complexity an issue. Their follow-up work [10] solved this problem by propagating terms though links and reducing the number of neighbors involved in the process.

In this work, we aim at better estimation of page authority based on the two assumptions proposed above. We 1) propose a model to propagate page authority based on the topicality concealed in the contexts among pages within diverse neighboring relationships; and 2) verify the superiority of our proposed model empirically by comparing it with many well-known algorithms.

2 Model

We extend traditional link analysis by allowing authority to flow among all possible neighboring pages, such that closely connected pages can directly influence the authorities of each other. The neighbor relationship is inferred from link structure in this work. Authorities from parent pages, children pages and sibling pages directly contribute to the current page. We also consider all pages on the web as a special type of neighboring pages of each individual since one is usually influenced by one's background. While the definition of neighboring pages isn't limited to the relationship inferred from link structure, we in this work focus on these four types of neighboring relationships described above. From another perspective, one page contributes its authority to all its neighbor pages on the web.

Once we determine where and how to distribute authority, the next step is to mine contexts between pages and use them to control authority flows. In traditional ranking models [2, 5], a parent equally splits its authority to all its children. We hypothesize that the authority flow distribution should prefer more similar neighbors, hence we incorporate a preference factor α to control the authority that flows to a specific type of neighbor. The preference factor α is determined based on the relative similarity between the current page to this type of neighbor. In particular, we simplify the representation of page topicality by using a taxonomy from Open Directory Project (ODP) [6] (we select its 12 top-level categories), and use a the Rainbow Naive Bayes text classifier [3], trained on 19,000 pages from each category. Thus each page can be represented by a 12-dimensional topical distribution, demonstrating the probability that the page belongs to each topic. We then use cosine similarity to represent how much the current page and its one neighbor are similar, which is given by: $rel(p_1, p_2) = \frac{\sum_i T_i(p_1)T_i(p_2)}{|T(p_1)||T(p_2)|}$, where

p_1 is p_2's neighbor, and $T(p_i)$ is the topical distribution of p_i. Since we distinguish the roles played by different types of neighbors, we use the average cosine similarity of the topicality between the current page and all neighbors within the same type to reflect how a page is similar to its specific type of neighbor. We use the centroid of topical distributions of all pages as the background distribution, and calculate the similarity between each page to this background as its similarity to the background neighbors. Specifically, we define the preference factor α with respect to each type as: $\alpha_i = \frac{avg\_rel(p_c,p_i)}{\sum_i avg\_rel(p_c,p_i)}$, where $i \in \{\text{Parent}(p_c), \text{Children}(p_c), \text{Sibling}(p_c), \text{Background}(p_c)\}$ and $avg\_rel(p_c, p_i)$ is the average similarity between the current page and the specified type of neighbor.

Once we determine the fraction of authority flowing to a specific type of neighbor, we next split the authority among all pages within the same type. To avoid an over-bias toward topical affinity, we distribute it uniformly within the neighbor type. When combining with the type-based preference factor, this step actually smoothes the authority propagation from the last step. We formalize the authority propagation model in Equation 1.

$$A^{(t+1)}(p_c) = \sum_i \alpha_i (\sum_{j:j \in i} \frac{A^{(t)}(p_j)}{O_i(p_j)}) \tag{1}$$

where $A(p_j)$ is the authority score and $O_i(p_j)$ is the number of links from the i^{th} type of neighbor to page p_j. This iterative process converges and results in a static probability distribution over all pages on the web. We then order pages by this score, and linearly combine the ranks with those from query-specific IR scores based on page content.

3 Experimental Results

Our goal is to use the proposed model to better estimate page authority and improve web search quality. Our dataset is the TREC (http://trec.nist.gov/) GOV collection, which is a 2002 crawl in .gov domain, containing around 1.25 million web pages. We used the topic distillation tasks in TREC web tracks in 2003 (50 queries) and 2004 (75 queries) to evaluate our approach. We compare our approach with Okapi BM25 [8], PageRank (PR), GlobalHITS (GHITS), Topical-Sensitive PageRank (TSPR) [2], and Topical PageRank (TPR) [5]. All these link-based ranking models combine with Okapi BM25 linearly by ranks. The parameters used in Okapi BM25 are set to be the same as Cai et al. [1]. We show the performance comparison at the best rank combination of query-specific scores and authority scores. We use precision@10, Rprec, MAP, and NDCG@10 to measure ranking quality.

Table 1 shows the performance comparison on TREC 2003 and 2004. On TREC 2003, our APR models outperform all other approaches on all metrics. The authority propagation from children to parents contribute more than that among siblings. But when combining all neighboring resources, the performance has significant improvement over the ones by propagating among only subsets of neighbor pages. Single-tailed student t-tests over P@10 and NDCG@10 show that the APR models significantly outperform PageRank and BM25 (p-value$<$0.01) at 95% confidence level.

On TREC 2004, our APR models show stable improvements over baselines. Even under the inconsistent performance on PageRank, the APR models show signifi-

Table 1: AuthorityPropagationRank and baselines performance.

	TREC 2003				TREC 2004			
	P@10	MAP	R-prec	NDCG@10	P@10	MAP	R-prec	NDCG@10
BM25	0.1200	0.1485	0.1398	0.1994	0.1907	0.1364	0.1769	0.2330
PR	0.1380	0.1538	0.1621	0.2197	0.2267	0.1523	0.1844	0.2790
GHITS	0.1360	0.1453	0.1574	0.2031	0.2147	0.1483	0.1758	0.2669
TSPR	0.1420	0.1718	0.1779	0.2414	0.1907	0.1429	0.1769	0.2390
TPR	0.1440	0.1691	0.1864	0.2301	0.2173	0.1532	0.1832	0.2714
APR(P+C+B)	0.1480	**0.1761**	0.1864	0.2281	0.2267	0.1661	0.1895	0.2874
APR(P+S+B)	0.1420	0.1518	0.1808	0.2127	0.2253	0.1557	0.1885	0.2806
APR(P+S+C+B)	**0.1540**	0.1752	**0.1922**	**0.2342**	**0.2333**	**0.1664**	**0.1941**	**0.2957**

cant improvement over BM25 (p-value$<$0.0001), PageRank (p-value$<$0.03), TPR (p-value$<$0.025), and PR (p-value$<$0.05) on P@10 and NDCG@10 at 95% confidence level.

4 Conclusion

We proposed an authority propagation model which propagates authorities among diverse types of neighboring pages based on topical affinity. Experimental results demonstrate its superiority on ranking performance over several representative link analysis algorithms. As future work, we expect to (1) generalize the concepts of neighboring nodes; (2) test the influence from the granularity of the taxonomy on ranking performance; and (3) investigate the sensitivity of the ways of interpreting the topicality context between pages with respect to ranking performance.

References

1. D. Cai, X. He, J.-R. Wen, and W.-Y. Ma. Block-level link analysis. In *Proc. 27th Annual Int'l ACM SIGIR Conf. on Research and Dev. in Information Retrieval*, July 2004.
2. T. H. Haveliwala. Topic-sensitive PageRank. In *Proc. of the 11th Int'l World Wide Web Conf.*, pages 517–526. ACM Press, May 2002.
3. A. K. McCallum. Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering. http://www.cs.cmu.edu/˜mccallum/bow, 1996.
4. L. Nie and B. D. Davison. Separate and inequal: Preserving heterogeneity in topical authority flows. In *Proc. 31st Annual Int'l ACM SIGIR Conf. on Research and Dev. in Information Retrieval*, pages 443–450, July 2008.
5. L. Nie, B. D. Davison, and X. Qi. Topical link analysis for web search. In *Proc. 29th Annual Int'l ACM SIGIR Conf. on Research & Dev. in Info. Retrieval*, pages 91–98, Aug. 2006.
6. The dmoz Open Directory Project (ODP), 2009. `http://www.dmoz.org/`.
7. T. Qin, T.-Y. Liu, X.-D. Zhang, Z. Chen, and W.-Y. Ma. A study of relevance propagation for web search. In *Proc. 28th Annual Int'l ACM SIGIR Conf. on Research and Dev. in Information Retrieval*, pages 408–415, 2005.
8. S. E. Robertson. Overview of the OKAPI projects. *Journal of Documentation*, 53:3–7, 1997.
9. A. Shakery and C. Zhai. A probabilistic relevance propagation model for hypertext retrieval. In *Proc. of the 15th ACM Int'l Conf. on Information and Knowledge Management (CIKM)*, pages 550–558, 2006.
10. A. Shakery and C. Zhai. Smoothing document language models with probabilistic term count propagation. *Inf. Retr.*, 11(2):139–164, 2008.