Knowledge-Based Systems 77 (2015) 56–67

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

On characterizing and computing the diversity of hyperlinks for anti-spamming page ranking

Bo Yang [a,b,*], Hechang Chen [a], Xuehua Zhao [a], Masato Naka [a], Jing Huang [a,b,*]

[a] College of Computer Science and Technology, Jilin University, China
[b] Key Laboratory of Symbolic Computation and Knowledge Engineering, Ministry of Education, China

ARTICLE INFO

Article history:
Received 16 November 2013
Received in revised form 26 December 2014
Accepted 28 December 2014
Available online 9 January 2015

Keywords:
Search engine
Page ranking
Hyperlink analysis
Probabilistic counting
Smart teleportation

ABSTRACT

With the advent of big data era, efficiently and effectively querying useful information on the Web, the largest heterogeneous data source in the world, is becoming increasingly challenging. Page ranking is an essential component of search engines because it determines the presentation sequence of the tens of millions of returned pages associated with a single query. It therefore plays a significant role in regulating the search quality and user experience for information retrieval. When measuring the authority of a web page, most methods focus on the quantity and the quality of the neighborhood pages that direct to it using inbound hyperlinks. However, these methods ignore the diversity of such neighborhood pages, which we believe is an important metric for objectively evaluating web page authority. In comparison with true authority pages that usually contain a large number of inbound hyperlinks from a wide variety of sources, it is difficult for fake authorities, which boost their page rank using techniques such as link farms, to occupy the high diversity of inbound hyperlinks due to prohibitively high costs. We propose a probabilistic counting-based method to quantitatively and efficiently compute the diversity of inbound hyperlinks. We then propose a novel link-based ranking algorithm, named Drank, to rank pages by simultaneously analyzing the quantity, quality and diversity of their inbound hyperlinks. The validations on both synthetic and real-world data show that Drank outperforms other state-of-the-art methods in terms of both finding high-quality pages and suppressing web spams.

1. Introduction

With the rapid development of novel web techniques, the amount of information on the Web has increased significantly. To retrieve useful information from this tremendous information resource, users rely on search engines that retrieve information that matches a users search query. The search results for a single query often contain tens of millions of entries, but users only make use of the top-listed ones [1]. Therefore, it is extremely important for all search engines to equip powerful page ranking algorithms that can arrange the most relevant entries as the top results in order to maximize the service quality and user experience.

By intentionally attacking the vulnerability of existing ranking algorithms [2,3], search engine optimizers are able to deceitfully boost the ranks of spam web pages. These spam web pages often contain low-quality and misleading or fake information, thereby decreasing the quality of a search engine's results. One of the main techniques for web spamming is to inject noises, such as spurious hyperlinks, into the Web. In response, the goal of anti-spamming is to design a robust ranking mechanism that is unaffected by introduced noises, enabling it to suppress malicious spam pages and preserve the normal ranks, i.e., non-spam ranks, of high-quality pages.

With the popularity of link-based page ranking, link-based spamming has become a common cheating method. Link-based spamming relies on special link structures constructed by artificially introduced hyperlinks to improve their ranks. For example, the construction of link farms using child node manipulation or link exchange has become a widely used spamming trick [4–6]. Compared with content-based anti-spamming methods [7,8], link-based methods perform much better in terms of both computation efficiency and spam suppression efficacy, thus attracting attention from both academic and industry. As search engine optimization techniques are continuously improved and developed upon, the study of anti-spamming will become a long-term task faced by information retrieval. We review recent developments in link-based ranking and anti-spamming before introducing our proposed method.

* Corresponding authors at: College of Computer Science and Technology, Jilin University, China.

http://dx.doi.org/10.1016/j.knosys.2014.12.028
0950-7051/

1.1. Related work

The primitive search engines were designed to satisfy the requirement of text retrieving based on content-based ranking, by considering the term frequency (TF) and the inverse document frequency (IDF) [9]. Link-based ranking became popular because content-based ranking was not scalable for web searching and the TF-IDF technique was easily cheated by content-based spamming techniques. The PageRank [10] and HITS [11] were two main link-based page ranking algorithms in the 1990s. According to PageRank, the greater number of times a page is directed to by other authoritative pages, the higher its rank will be. Armed with PageRank, Google became the most successful commercial search engine [12–15]. According to HITS, there are two types of pages, authority and hub pages. Authority pages are expected to provide the information that users need; hub pages are expected to contain many hyperlinks pointing to authority pages. In the past 15 years, many improvements have been made to these two basic link-based ranking mechanisms. For examples, to solve the problem of "topic drift" faced by HITS, the ARC (Automatic Resource Compilation) [16] was proposed to assign different weights to hyperlinks according to anchor texts and query keywords in context; to decrease the time complexity of ranking, the SALSA (Stochastic Approach for Link Structure Analysis) [17] introduced a mechanism to integrate a random walk model to HITS. To increase the relevancy between queries and answers, the Hilltop [18] and the TSPR (Topic-Sensitive PageRank) [19] were proposed to enhance PageRank by either suppressing the effect of spamming or by utilizing the semantics provided by ODP (open directory project). The SD (Similarity Down-weighting) and the SC (Sequential Clustering) [20] were proposed to address the interference of TKC (tightly knit communities), an important issue faced by link-based ranking.

With the development of link-based page ranking, link-based web spamming is becoming more and more common. In suppressing artificially introduced spam hyperlinks, the concept of "trust" and the model of "trust diffusion" were introduced into the process of page ranking. Gyangyi et al. raised the point that high quality pages are rarely linked to grunge pages and proposed the TrustRank algorithm [21]. The TrustRank algorithm selects trustworthy authority pages as seeds by assigning high initial authority values to them, and then adopts the same iteration process as PageRank to simulate the diffusion of trust values along hyperlinks. Following a similar idea, many improvements were proposed including Topical TrustRank [22], Anti-TrustRank [23], BadRank [24]. Interestingly, Anti-TrustRank is actually an inverse version of TrustRank, which is based on the point that grunge pages are rarely linked by high quality pages. Furthermore, Wu et al. suggested combining the two strategies of TrustRank and Anti-TrustRank to determine whether or not a page is spam [25]. Very recently, some ideas from physics were borrowed to design page ranking algorithms to further suppress cheating behaviors of spam hyperlinks. For example, the DiffusionRank [26] algorithm regards the authority of pages as heat and hyperlinks as tubes transferring heat so as to simulate the authority diffusion through pages in a way similar to how heat is spread. Distinctly, the AIR (Affinity Index Ranking) [27] regards the Web as a circuit in which the hyperlinks, page authorities and authority diffusion are modeled by diodes, voltages and electricity moving through a circuit. Since diodes conduct electricity in a single direction, page authority is also expected to flow in a single direction from high-voltage pages to low-voltage pages, thereby suppressing the boosted effect of link farms consisting of a group of maliciously introduced and densely connected low-quality pages.

1.2. Motivation and contribution

When evaluating the authority of pages, the existing link-based ranking algorithms focus on the quantity and quality of inbound hyperlinks. In this study, we consider a third factor, the diversity or the heterogeneity of inbound hyperlinks. In the Web, different pages (or sites) actually represent different parties with respective interests, and hyperlinks are constructed deliberately, rather than randomly, and can be thought of metaphorically. Normally, a hyperlink from A to B is constructed in a way that A recognizes B. Intuitively, the evaluation of page authority is akin to a voting game, in which hyperlinks denote votes, and the quantity, the quality and the diversity of hyperlinks correspond to the number of votes, the weight of voters, and the diversity of parties voters represent, respectively. If a web page accepts a large number of votes (i.e. inbound hyperlinks) from a variety of parties, it implies that the page has a widely recognized reputation and is deemed a veritable authority. On the other hand, if most of inbound hyperlinks of a page come from a very limited range of parties, even though the amount of such links is large, it may not qualify as a high authority, and is likely suspected of spam page cheating. The new metric of diversity will hopefully facilitate the development of a better ranking algorithm in which high-quality pages, low-quality pages and spam pages are expected to be boosted, suppressed and discriminated, respectively. Since it is economically expensive to construct spam organizations, search engine optimizers mainly focus attacks on commercial search engines possessing the largest market shares [28]. In this study, we work to improve the performance and robustness of the PageRank framework, which is the foundation of Google's ranking system, by introducing the concept of diversity. Based on the novel authority metric, we then contribute a new random walk-based ranking framework, referred to as Drank (diversity based page rank), in which the quantity, quality, and diversity of inbound hyperlinks are comprehensively explored to evaluate page authority.

Previously, we have briefly present and preliminarily validate the aforementioned idea [29], which will be extensively extended in this work by supplementing new models and algorithms, such as the diversity computation based on URL analysis and the authority computation based on smart teleportation, and meanwhile sufficient experimental validations. The rest of this paper is organized as follows. In Section 2, we introduce the concept of diversity, the scalable computation of diversity, the computation of diversity-specific authority, and the framework of Drank. In Section 3, we conduct a series of experiments to rigorously validate the efficacy of the Drank by comparing it with the state-of-the-art anti-spamming methods on both synthetic and real-world data. Finally, Section 4 concludes the work by highlighting its main contributions.

2. Drank algorithm

2.1. The framework of Drank

Spam and normal pages are different in terms of their hyperlink structures. The inbound hyperlinks of normal pages come from a variety of sources, whereas those of spam pages mainly come from some specific sources, e.g. link farms, in which hyperlinks are significantly dense and pages are seldom linked by outside world (Fig. 1). This distinction can be exploited by anti-spamming page ranking.

Currently, the primary web spamming technique is to construct link farms in terms of child node manipulation or link exchange. Formally, a link farm is defined as follows.