Knowledge-Based Systems 70 (2014) 118–127

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Short Communication

Approximating web communities using subspace decomposition

Justine Eustace, Xingyuan Wang *, Junqiu Li

Faculty of Electronic Information and Electrical Engineering, Dalian University of Technology, Dalian 116024, China

ARTICLE INFO

Article history:
Received 22 October 2013
Received in revised form 10 June 2014
Accepted 11 June 2014
Available online 19 June 2014

Keywords:
Web communities
Web graphs
Subspace decomposition
Information retrieval
Community detection
Spam detection

ABSTRACT

Herein, we propose an algorithm to approximate web communities from the topic related web pages. The approximation is achieved by subspace factorization of the topic related web pages. The factorization process reveals existing association between web pages such that the closely related web pages are extracted. We vary the approximation values to identify varied degrees of relationship between web pages. Experiments on real data sets show that the proposed algorithm reduces the impact of unrelated links and therefore can be used to control spam links in web pages.

1. Introduction

Degree centrality is a common characteristic used in community discovery. It identifies and measures participation of edges and vertices in a community structure [2]. Degree centrality is also used to detect intermediate edges between communities [14]. In the World Wide Web, degree centrality is used to identify the most informative web pages. When degree centrality is used to rank web pages, the existence of noise (spam) links can influence the ranking score of web pages. Generally these noisy links obscure the distinguishing characteristics of topic related web community [33]. The presence of noisy links can further cause topic drift, a compromised on information security system as well as wastage of storage space.

In this paper, we propose an algorithm which exploits the topological relationship in web pages. In the proposed algorithm, first a set of web-links are extracted from retrieved web pages and then expanded by extracting their in-links and out-links. Each link is verified by using subspace decomposition. The process of verification reveals the links relationships between the web pages. Links with higher degree of correlation are retained. Finally, we apply PageRank and HITS algorithms on retained links.

Given a set of retrieved web pages the PageRank algorithm [6] create its corresponding topic related web community by ranking pages based on their link structure. Under the PageRank algorithm ranking is done by assigning global static values to web pages which are then propagated to other linked web pages. Similarly, the HITS algorithm [21] ranks topic related web pages and creates a bipartite community of hub and authority cores. It ranks related web pages by computing corresponding authority and hub scores of each linked web page. In summary, the contributions of this paper are as follows;

- We theoretically described the subspace approximation methods used in this paper (SVD and ULVD) and their application in community detection.
- We proposed a framework which uses the subspace approximation method to approximate a community of query related web pages. Our proposed method, considers the influence of noise-links on the quality of query related web pages.
- We further introduced the use of ULVD in community detection. The experimental results indicate that ULVD out performs SVD-based algorithms in some cases.
- We evaluated our framework on a real data set. Using HITS and PageRank algorithms, we evaluated the detected web communities. Moreover, empirical results suggest that the proposed method succeeds in removing noise-links in a community of query related web pages.

The next six sections, we organize this paper as follows; in Section 2, we give a brief overview of related work. In Section 3, we introduce key concepts used in this paper. We briefly discuss types of subspace decompositions which are used in this work in Section 3.3, where brief overview of SVD and ULVD decompositions is given. In Section 4, we discuss the proposed algorithm in fair

* Corresponding author. Tel.: +86 1307 4102070.
E-mail addresses: justineustace@yahoo.com (J. Eustace), wangxy@dlut.edu.cn (X. Wang), meiliqiutian@126.com (J. Li).

http://dx.doi.org/10.1016/j.knosys.2014.06.017
0950-7051/© 2014 Elsevier B.V. All rights reserved.

detail. From Sections 5–5.4, we present the setup of experiments and offer a thorough discussion of the results. We conclude this work and offer a brief overview of possible future work in Section 6.

2. Related works

Several algorithms have been proposed as mechanisms to improve the quality of identifying web communities [40]. These include heuristic methods, such as, the submission of specific query terms to the search portal or a combination of document analysis and ranking algorithms.

Other methods include the use of weighted ranking methods among others [24,42,31,23,25]. Recently Benzi et al. [3] have extended the concept of sub-graph centrality to rank hubs and authorities in a web community. However, their proposed method does not consider the effect of noisy links.

These methods can be classified into content-based methods, link-based methods, and heuristic-based methods [33]. Content-based methods include the earliest methods like [12,28] and recently [26,38,11,30]. These methods uses statistical verification techniques to identify noise- links.

Link-based methods include [17,36,27,9,15], see [33] for further survey details and use properties of the web links in a network (such as topological, distinct node properties and link labels).

Heuristic based methods such as [41,23,25] use a combination of existing techniques (link-based, content-based, machine learning or other methods like recommendation systems). Recommendation systems [35,39,8,32], harvest user search experiences and suggest appropriate search results. However, recommendation systems, are often influenced by noise-links as well (depending on the algorithm that is implemented).

To counter noise-links in recommender systems, Wang et al. [37], proposed a link-based method which is a modified version of the PageRank algorithm. Similarly, Champin et al. [8] proposed a link-based method to improve search results in HeyStaks[1] recommendation system. The technique shows satisfactory results in countering spam, however, the precise contribution to the score of authoritative webpages, is unknown.

Related to our work is, Noise Page Elimination Algorithm (NPEA) by Hou and Zhang [19]. In their algorithm, a quality web-community is constructed by removing unrelated web-links. Although their approach appears strong in theory, our empirical investigation (as shown in Figs. 3 and 4 with further details in Tables 2–5 of Appendix A) reveals that it fails to retain most of the query-related links, in most cases. Unlike NPEA algorithm which retrieve a web community using a projected vector of only query retrieved web pages [19]. The proposed algorithm, retrieve a web community of query related links from a subspace of all links related to query and their corresponding web pages.

3. Preliminaries

3.1. The web as a graph

A web graph is a graph $G = (V, E)$ where V, is a non-empty set of web pages (vertices) and E is the set of hyperlinks between web pages (edges). Each edge ($e_{ij} \in E$) presents a directed connection of ordered pair $(v_i, v_j) \in V$, such that web page $v_i \in V$ is linked to a web page $v_j \in V$. $|V|$ is the total number of web pages in the web graph. The degree of vertex v_i i.e., $\delta\ (v_i)$ is defined as the number of edges linked to web page v_i. A non-empty web graph $G(V, E)$ is a directed graph where $e_{ij} \in\ E \nRightarrow e_{ji} \in E$.

In a directed graph, the number of edges of which a vertex v_i is directed to is called *out-degree*. *In- degree* is the number of edges which are directed to vertex v_i. An alternating sequence of vertices, linked by consecutive edges $v_1, e_{12}, v_2, e_{23}, \ldots, v_n, e_{nn+1}, v_{n+1}$ where $n > 0$ and $e_{ii+1|1\leqslant i\leqslant n}$ is called a *walk*. A path is a walk in which its sequence of vertices is distinct. Two web pages $v_i,\ v_j \in V$ are said to be connected in G, if there exists a path from v_i to v_j. The maximal connected sub graph $C \in G$ is therefore a connected component.

Adjacency matrix $\mathbf{A}$, is used to realize the graph G, where $\mathbf{A}_{ij} = 1$ if there is an edge between vertex v_i and vertex v_j; $\mathbf{A}_{ij} = 0$ otherwise. Note that, self-loops and parallel edges are ignored. For a vertex pair $(v_i, v_j) \in V$, a link from v_i to v_j is considered as an in-link if there exists an edge, $e_{ij} \in E$ where $\mathbf{A}_{ij} = 1$. A link is considered as an out-link when there exists an edge e_{ji} where $\mathbf{A}_{ji} = 1$.

Generally, a community is a sub-graph $C \in G$ which is densely interconnected and sparsely connected to the rest of the graph. In the World Wide Web, a typical community can be a set of topic related web pages. Related web pages in a web community can be grouped into authoritative cores. These cores can be identified using centrality measures, and ranked based on their authoritative score.

3.2. Link based ranking algorithms

3.2.1. Hyperlink-Induced Topic Search (HITS)

HITS is a link based algorithm which relies on references within web pages for ranking [21]. These references are categorized into hubs and authorities web pages. Authority web pages contain relevant information, while hub web pages simply refer to authority web page. Both hubs and authorities web pages influence each other [21] i.e., a good authoritative web page is pointed by good hubs.

Let $\mathbf{h}, \mathbf{a} \in \mathbb{R}^{1\times m}$ represent hub and authority weight vectors of m web pages, respectively. For a web page v_i, $\mathbf{h}(v_i)$ and $\mathbf{a}(v_i)$ are the hub and authority weights of v_i, respectively. Initially hub and authority weights are assigned non-negative values, and updated iteratively until they converge. Let $t = 1,2,3,\ldots,\infty$ be the number of iterations, such that $\mathbf{h}$ and $\mathbf{a}$ vectors are updated as:

$$\begin{aligned} \mathbf{h}(v_i)^t &= \sum_{v_i \mapsto v_j} \mathbf{a}(v_j)^{t-1} \quad \text{and} \\ \mathbf{a}(v_i)^t &= \sum_{v_j \mapsto v_i} \mathbf{h}(v_i)^{t-1} \quad \text{where } v_i \mapsto v_j, \text{ means } v_i \text{ is linked with } v_j \end{aligned} \tag{1}$$

To avoid entry value divergence, $\mathbf{h}$ and $\mathbf{a}$ vectors are normalized, such that their sum on each entry is equal to 1. For m linked web pages with adjacency matrix $\mathbf{A}$, where ϵ_1 and ϵ_2 are normalization factors, Eq. (1) then becomes:

$$\mathbf{h}^t = \epsilon_2 \mathbf{A}\mathbf{A}^T \mathbf{h}^{t-1} \quad \text{and} \quad \mathbf{a}^t = \epsilon_1 \mathbf{A}^T \mathbf{A} \mathbf{h}^{t-1} \tag{2}$$

The dominant eigenvectors of $\mathbf{AA}^T$ and $\mathbf{A}^T\mathbf{A}$ are respectively the right and left singular values of $\mathbf{A}$. These eigenvectors are corresponding hub and authority scores of m linked web pages.

3.2.2. PageRank algorithm

PageRank is a query independent ranking algorithm which ranks web pages based on their topological structure [6]. The authority scores are computed using weight values which are calculated from their in-linked web pages. Relying on the assumption, good web pages are linked by good authoritative web pages, authoritative web pages carry more weight compared to the less authoritative ones.

PageRank algorithm is described as in [6,7,22]. Suppose a web page p_j is linked to m web pages $(p_1, p_2, \ldots, p_m)$, where $\mathbf{A}$ is its corresponding adjacency matrix such that $\mathbf{A}_{ij} = 1$, if p_i is linked with p_j,

[1] http://www.heystaks.com.