Exploration of Web Page Structural Patterns Based on Request Dependency Graph Decomposition

Cheng Fang, Beijing University of Post and Telecommunications, Beijing, China
Bo Ya Liu, Beijing University of Post and Telecommunications, Beijing, China

This article first proposed a Bipartite Request Dependency Graph (BRDG) that describes the object-level interrelationships between user click requests and embedded web object requests. These two kinds of requests are classified from HTTP data by an identification algorithm. The interrelationships between user click requests and embedded web object reflect the web page structural, which contain latent web information. Exploring structural patterns is crucial for many aspects like web security analysis and web information visualization. Accordingly, the article also proposed a novel graph decomposition method called orthogonal nonnegative matrix tri-factorization (tNMF) to the BRDG. Compared to traditional web graph analysis focus on statistical and structural properties of the whole graph, the proposed method is dedicated to mine latent web structural patterns. Decomposition results demonstrate that several interesting structures exist in the BRDG. The article aims at classifying these subgraphs as several structural patterns and shedding light on the causes of these patterns.

KEYWORDS
Bipartite Request Dependency Graph (BRDG), Graph Decomposition, Internet, Web Information Visualization, Web Page Structural

INTRODUCTION

With rapid development of next generation networks (NGN) proposed by (ITU-T, 2006), heterogeneity and complexity in networks have become increasingly prominent. In the meanwhile, to manage and secure the networks, information visualization is imperative for administrators. Under these circumstances, many researchers have devoted themselves to web usage mining and obtained numerous achievements. This paper focuses on exploring latent interrelationships between web objects on World Wide Web (WWW).

One of the most widely used model for web structural analysis is Web graph, where the nodes are web pages and the edges are the hyperlinks between them. (Broder, 2000) conducted the first large-scale study on the Web graph and presented a bow-tie picture consisting of three distinct components in almost equal size to describe the macroscopic structure of the web. On this basis, (Donato, 2005) offered a better understanding of inner structures in the Web graph. Besides, (Huang & Lai, 2003) proposed a new approach to cluster the Web graph and applied it to web visualization. Nevertheless, Web graphs only represent the relationship between hyperlinks, without considering the context. (Sethu & Yates, 2010) achieved hyperlink classification using text mining analysis and built a multi-relational web graph. As compared
with the fixed web, the mobile web is structurally different. (Jindal, 2008) found that the connectivity of mobile web was sparser than the fixed one and the node degree distributions fell off much more rapidly. (Liu & Ansari, 2014a) identified website communities successfully in mobile Internet based on affinity measurement.

This paper introduces the notion of the Bipartite Request Dependency Graph (BRDG), a graph derived from a dependency graph model in (Liu, 2014b), which took a two-step algorithm to identify user clicks from a plenty of HTTP requests. In other words, the BRDG is a useful application of the dependency graph model. The authors’ study finds that the BRDG is very large, sparse and seemingly complex. To explore latent web structural patterns, the authors apply a tNMF-based graph decomposition method to the BRDG and extract a number of interesting structural properties. The tNMF is a co-clustering algorithm, which has been shown to be useful in many applications, such as identifying suspicious activities through DNS failure graph in (Jiang, 2010), and understanding the intensive and continuous data usage patterns from mobile users (Jin, 2012).

The major contributions of this paper lie in three respects: (i) the authors propose the bipartite dependency graph to describe the interrelationships between user click requests and embedded web object requests in mobile Internet; (ii) the authors implement the tNMF-based decomposition method on BRDG; (iii) based on decomposition results, the authors classify them as five structural patterns and reveal the causes of these patterns. The rest of this paper is organized as follows. Section 2 introduces related work. Section 3 mainly introduces the definition and overall characteristics of the BRDGs. Section 4 presents the graph decomposition and classification methodology. Section 5 summarizes decomposition results as five structural patterns and gives meaningful interpretations. Section 6 concludes our paper briefly and proposes the future work.

RELATED WORK

To reduce user perceived latency, some researchers have concentrated on studying web objects, which are the basic elements of a web page, to achieve web prefetching. The dependency graph (DG) (Padmanabhan & Mogul, 1996) first proposed by Padmanabhan and Mogul depicts the pattern of accesses to the web objects for web prefetching. Based on the DG, (Domenech, 2006) presented an improved graph model called the double dependency graph (DDG), which distinguishes two classes of dependency between objects of the same page and objects of different pages. Another graph model to describe web object dependencies is Referrer Graph (RG) (de la Ossa, 2010). The advantage of the RG lies in high efficiency by reducing the number of edges in the graph to save computational and memory resources.

BRDG of this article is derived from a request dependency graph model proposed in (Liu, 2014b). When a user clicks a new Web page, the Web client sends a HTTP request to Web server which the article defines as the user click request. The Web server returns the initial page usually containing a number of hyperlinks of the embedded objects. After parsing these hyperlinks, the Web client sends a set of requests to get those embedded objects, which the article defines as embedded requests. All user click requests and embedded requests are identified by Universal Resource Locators (URLs). The request dependency graph (RDG) is used to model the dependency relationships among web objects. The request dependency graph is defined as a directed and weighted graph, \( G=(O,S,E,W) \). \( O \) is the set of nodes representing accessed objects which are identified by URL requests. Each node is assigned an occurrence count, which compose set \( S \). \( E \) is the set of directed edges with weights \( W \). An edge exists from node \( O_i \) to \( O_j \) when they meet two conditions: (1) for a request sequence \( v = \{r_1, r_2, \ldots r_j\} \)
Internet Crime: How Vulnerable Are You? Do Gender, Social Influence and Education play a Role in Vulnerability?
www.igi-global.com/chapter/internet-crime-vulnerable-you-gender/60937?camid=4v1a

www.igi-global.com/chapter/model-charter-conclusion/55538?camid=4v1a