Towards a Formal Concept Analysis Approach to Exploring Communities on the World Wide Web

Jayson E. Rome and Robert M. Haralick

Department of Computer Science,
The City University of New York, New York NY 10016, USA

Abstract. An interesting problem associated with the World Wide Web (Web) is the definition and delineation of so called Web communities. The Web can be characterized as a directed graph whose nodes represent Web pages and whose edges represent hyperlinks. An authority is a page that is linked to by high quality hubs, while a hub is a page that links to high quality authorities. A Web community is a highly interconnected aggregate of hubs and authorities. We define a community core to be a maximally connected bipartite subgraph of the Web graph.

We observe that the web subgraph can be viewed as a formal context and that web communities can be modeled by formal concepts. Additionally, the notions of hub and authority are captured by the extent and intent, respectively, of a concept. Though Formal Concept Analysis (FCA) has previously been applied to the Web, none of the FCA based approaches that we are aware of consider the link structure of the Web pages. We utilize notions from FCA to explore the community structure of the Web graph. We discuss the problem of utilizing this structure to locate and organize communities in the form of a knowledge base built from the resulting concept lattice and discuss methods to reduce the complexity of the knowledge base by coalescing similar Web communities. We present preliminary experimental results obtained from real Web data that demonstrate the usefulness of FCA for improving Web search.

1 Introduction

Traditional techniques for information retrieval involve text based search and various indexing methods. The presence of hyperlinks between documents presents challenges and opportunities that traditional information retrieval techniques have not had to deal with. By viewing the set of n pages on the World Wide Web as nodes V and links (similarity, association) between pages as directed edges E of a directed graph $\Gamma = (V, E)$, the graph of n nodes can be stored in an $n \times n$ matrix. A nonzero entry in the $(i, j)^{th}$ position of the matrix indicates an edge (possibly weighted or labelled) from node i to node j. A hyperlink implies some form of endorsement, or conferral of authority, by citing document to the cited document. A large portion of the current research in improving web search

B. Ganter and R. Godin (Eds.): ICFCA 2005, LNAI 3403, pp. 33–48, 2005.

is concerned with utilizing the hyperlinked nature of the web. Kleinberg's HITS algorithm [1], and various extensions [2], [3], [4], [5], and the Google PageRank algorithm [6], [7] demonstrate the success of link based ranking in refining Web search. Henziger's recent survey [8] enumerates the following open algorithmic challenges for Web search engines:

- Finding techniques to generate random samples of the Web in order to determine statistical properties of the Web,
- Modeling the web to explain observed properties,
- Detecting duplicates and near duplicates to improve search efficiency,
- Analyzing temporal trends in data streams that result from user access logs,
- Finding and analyzing dense bipartite subgraphs, or Web communities,
- Finding eigenvector-induced partitionings of directed graphs in order to cluster the Web graph.

1.1 Hubs and Authorities

Consider the problem of finding "definitive" or "authoritative" sources in the mass of information available on the web. The user should be provided with relevant pages of the highest quality. The hyperlink structure of the web contains a tremendous amount of latent information in that the creation of a link from page a to page b in some way represents a's endorsement of b. Purely text based search methods fail to find authoritative sources. For example if one uses the query "operating systems," there is no guarantee that Windows, Linux, Apple or any other operating system vendor will be among the pages returned because these pages may not explicitly contain the query terms. These pages are, however, relevant and of high quality and should in fact be returned. We can define these pages to be *authorities* because they are linked to by a large number of other pages. We can define a *hub* to be a page with a large collection of links to related pages. A good hub should point to many good authorities and a good authority should be pointed to by good hubs [1].

1.2 Web Communities

An interesting problem associated with the Web is the definition and delineation of so called Web communities [9], [10], [11], [12], [13], [14]. A *web community* is loosely defined to be a collection of content creators that share a common interest or topic and manifests itself as a highly interconnected aggregate or subgraph [9]. Kumar et al define a web community as being "characterized by dense directed bipartite subgraphs [10]." Figure 1 illustrates a simple community centered around a densely interconnected set of hubs and authorities. The World Wide Web contains many thousand explicitly defined communities and many more that are implicitly defined or are emerging [10]. The systematic extraction of emerging communities is useful for many reasons, including communities provide high quality information to interested users, they represent the sociology of the web and they can be used for target advertising [9]. In addition, community linkage can be used to find association between seemingly unconnected topics.