Reading Report - Week 3 **Varad Meru**
CS 221 - Information Retrieval Student # 26648958
Prof. Cristina Lopes Due Date: 01/24/2015

The Web as a graph*: A Summary

By Ravi Kumar, et al.

Summary

The authors start with the description of the *Web graph*, where the nodes represent static html pages and directed edges are the links formed by the hyperlinks present in the page. On an average, there are seven hyperlinks present in a page, with the number of nodes estimated to be over a billion. This humongous structure would require specialized algorithms, and metrics to be understood and models to describe the nature of the problem, which is covered in various sections of the paper. This paper also highlights the relevant work done by many of the researchers in the field of graph theory, databases and information retrieval.

The algorithms discussed in the paper were HITS method and extensions, the enumeration of certain bipartite cliques and classification algorithms using hyperlinks. The HITS method focuses on classifying the pages as hubs and authorities based on numerical values assigned to the authority and the hub weights. The authority weights are computed by adding all the hub weights of the incoming page links. Similarly, the hub weight is computed by adding the authority weights of the pages to which the link is going to. The final values would be independent of the initial values and are computed by the *intrinsic* features of the collection. Extensions of the HITS algorithm work on the update of the authority and the hub scores in a non-boolean manner. The Topic enumeration problem was focused on presenting a snapshot of all the communities given the snapshot of the web. A *bipartite-core* $C_{i,j}$ is defined to be a graph on $i + j$ nodes that contains at least one $K_{i,j}$ as a subgraph. Web graph would be suggestive of a "cyber-community". The process of *trawling* is to find all the A *bipartite-cores*. The *elimination-generation paradigm*. An algorithm in this paradigm performs number of sequential passes over web-graph, which is stored as a binary relation. *Elimination* refers to the task of removing some of the nodes which would not fulfill some conditions. *Generation* refers to pruning of nodes which just make the cut after the elimination phase. It is continued until there is nothing substantial to be done. The final algorithm focuses on supervised learning and classifying using pre-defined categories. To accomplish the same, statistical methods like Markov random fields and *relaxation labeling* technique for better categorization may be used.

The degree distributions is checked for each node. It is seen that the fraction of web pages with in-degree i is proportional to $1/i^x$ for some $x > 1$. The standard value of x looks to be $x = 2.1$. The best fit line in this experiment gives a power law with $x = 2.72$. The average out-degree is about 7.2. The enumeration of bipartite experiment found well over hundred thousand cores of values of i, j in the range 3-4. From the connected component analysis, a *weakly-connected component* is a set of pages each of which is reachable from any other if hyperlink may be followed either forward or backwards. The largest weakly-connected component has 186 million nodes. A *strongly-connected component* is a set of pages such that all pairs of pages (u, v) in the set, there exists a directed path from u to v. The largest strongly-connected component has roughly 56 million nodes. The paper also describes the models of graphs.

Commentary

The paper presents the idea of using graphical methods as well as algorithms well on very large graphs. The application of filtering and elimination can be seen in simplest of algorithms as well as in the cases where the context needs to be derived. It also described the efficiency of the algorithms with the measurements and the graphical models presented.

*Kumar, Ravi, et al. *The Web as a graph.* Proceedings of the nineteenth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems. ACM, 2000.