Chapter X

Data Cleansing and Validation for Multiple Site Link Structure Analysis

Mike Thelwall
University of Wolverhampton, UK

ABSTRACT

A range of techniques is described for cleansing and validating link data for use in different types of Web structure mining, and some applications are given. The main application area is Multiple Site Link Structure Analysis, which typically involves mining patterns from themed collections of Websites. The importance of data cleansing and validation stems from the fact that Web data are typically very messy. It involves extensive duplication of pages and page components, which when analyzing raw Web data may give meaningless results.

INTRODUCTION

Hyperlinks are important for their use in navigation within a site, and Web usage mining algorithms have been built to cluster site users based upon navigation patterns (Hafri et al., 2003) and to create site link structures from Web server logs (Wu & Ng, 2003). Links are also used for directing visitors to other sites, however, and links on other sites similarly provide a potential source of new visitors. Inter-site links, the focus of this chapter, are now routinely exploited by Web applications for a variety of purposes,
including building search engine indexes, ranking search results, identifying clusters of topic-related pages, and extracting information about the behavior of Web authors. Such is the importance of links that there is a significant body of work that chooses them as the objects of study, for modeling their creation or for analyzing the structures produced by them (Barabási, 2002).

This chapter covers two different Web structure mining approaches, *link topology mining* and *link URL mining*, which use different raw data and methods. Link topology mining treats the Web as a graph with the pages as nodes and the links as either edges (directionless) or arcs (preserving the link direction), discarding all information about the content of the pages themselves. For link URL mining, the topology is used in conjunction with the URLs of the source and target of each page, allowing the mining of more concrete linking patterns. Typical outputs include information about the interconnectivity of sets of Websites or their linking practices. Both types of link mining can be used either on their own or in conjunction with other approaches to solve specific problems such as Web topic identification.

The techniques described in this chapter are primarily useful for a new type of Web structure mining, *Multiple Site Link Structure Analysis*, although they also have potential uses for Web Information Retrieval (IR). Web IR applications of link URL mining and link topology mining include heuristics to rank pages or cluster them by connectivity patterns, as well as to identify site types from their link structures alone. Multiple Site Link Structure Analysis (MSLSA) is the analysis of the link structure of a *collection* of Websites. The emphasis is on a finite collection of sites: more than one site but not the whole Web. The collection would typically have a unifying theme such as UK universities or U.S. libraries. The following are illustrative of the types of problems that might trigger an MSLSA investigation.

- What are the patterns of interlinking and clustering of the Websites of different nations/industry sectors/groups of countries?
- Which countries/types of site/sites do the Websites within a given nation/industry/sector/group of companies tend to link to?
- Which Websites out of a given set appear to be having the most (link based) impact on the Web?

These questions can be useful for a variety of different interest groups, including Website managers. On a larger scale this includes managers with the responsibility to ensure that collections of sites are effective, from the level of the sites of a large company or university to the national level. The techniques are also being investigated on an international scale for the European Union ([www.webindicators.org](http://www.webindicators.org)).

The objective of this chapter is to present a state-of-the-art summary of data cleansing and validation techniques for link topology mining and link URL mining, mainly for MSLSA but also for Web IR. A range of types of application techniques will also be given. The choice of cleansing and validation as the focus of the chapter is due to the problematic nature of data collection and interpretation on the Web. Data cleansing is important because of the many types of problems in raw Web data, including the existence of duplicate sites. Data validation is critical for MSLSA (but not Web IR) in order to be able to interpret findings. Current research suggests that the statistical techniques that can successfully analyze link structure data will vary by the scope of the study and so it will not be fruitful to go into the details of individual case studies here.