A Clustering-Based Approach for Weblog Data Cleaning

Authors : Amine Ganibardi, Cherif Arab Ali

Abstract : This paper addresses the data cleaning issue as a part of web usage data preprocessing within the scope of Web Usage Mining. Weblog data recorded by web servers within log files reflect usage activity, i.e., End-users' clicks and underlying user-agents' hits. As Web Usage Mining is interested in End-users' behavior, user-agents' hits are referred to as noise to be cleaned-off before mining. Filtering hits from clicks is not trivial for two reasons, i.e., a server records requests interlaced in sequential order regardless of their source or type, website resources may be set up as requestable interchangeably by end-users and user-agents. The current methods are content-centric based on filtering heuristics of relevant/irrelevant items in terms of some cleaning attributes, i.e., website's resources filetype extensions, website's resources pointed by hyperlinks/URIs, http methods, user-agents, etc. These methods need exhaustive extra-weblog data and prior knowledge on the relevant and/or irrelevant items to be assumed as clicks or hits within the filtering heuristics. Such methods are not appropriate for dynamic/responsive Web for three reasons, i.e., resources may be set up to as clickable by end-users regardless of their type, website's resources are indexed by frame names without filetype extensions, web contents are generated and cancelled differently from an end-user to another. In order to overcome these constraints, a clustering-based cleaning method centered on the logging structure is proposed. This method focuses on the statistical properties of the logging structure at the requested and referring resources attributes levels. It is insensitive to logging content and does not need extra-weblog data. The used statistical property takes on the structure of the generated logging feature by webpage requests in terms of clicks and hits. Since a webpage consists of its single URI and several components, these feature results in a single click to multiple hits ratio in terms of the requested and referring resources. Thus, the clustering-based method is meant to identify two clusters based on the application of the appropriate distance to the frequency matrix of the requested and referring resources levels. As the ratio clicks to hits is single to multiple, the clicks' cluster is the smallest one in requests number. Hierarchical Agglomerative Clustering based on a pairwise distance (Gower) and average linkage has been applied to four logfiles of dynamic/responsive websites whose click to hits ratio range from 1/2 to 1/15. The optimal clustering set on the basis of average linkage and maximum inter-cluster inertia results always in two clusters. The evaluation of the smallest cluster referred to as clicks cluster under the terms of confusion matrix indicators results in 97% of true positive rate. The content-centric cleaning methods, i.e., conventional and advanced cleaning, resulted in a lower rate 91%. Thus, the proposed clustering-based cleaning outperforms the content-centric methods within dynamic and responsive web design without the need of any extra-weblog. Such an improvement in cleaning quality is likely to refine dependent analysis.

Keywords : clustering approach, data cleaning, data preprocessing, weblog data, web usage data
Conference Title : ICDM 2018 : International Conference on Data Mining
Conference Location : Dublin, Ireland
Conference Dates : July 23-24, 2018

Open Science Index, Computer and Systems Engineering Vol:12, No:07, 2018 publications.waset.org/abstracts/93198.pdf