文献信息

- 名称：Web Usage Mining: A Review of Recent Works
- 作者：Rosli Omar, Abu Osman Md Tap, Zainatul Shima Abdullah
- 期刊：Information and Communication Technology for The Muslim World (ICT4M)
- 时间：2014
- 片断出处：INTRODUCTION
- 字数：545

Cooley, Mobasher and Srivastava [2] have further categorized web mining into three main components: web usage mining(WUM), web structure mining(WSM) and web content mining(WCM). WCM is the task of discovering useful information based on the content of web pages. Web contents include multimedia data, structured content such as XML documents, semi-structured such as HTML documents and unstructured data such plain text. Web content mining applications include the task of organizing and clustering the web pages based on content and as well as ranking of web pages basedtion on contents. WSM focuses on the structure of web sites using source data in the form of the structural information present in Web pages; typical applications are link-based categorization of Web pages, ranking of Web pages through a combination of content and structure and reverse engineering of Web site models. Taxanomy of WUM literature is presented in Figure 1.

Due to enormous interest in this field, there are plenty of studies being done since the last decade. In addition [expand: entertain: booster], there are a number of existing works on reviewing web mining and WUM approaches by focusing issues at the different levels of WUM namely pre-processing, pattern discovery and pattern analysis as in [4][5][6][7][8][9]and [10]. Chitraa [4] and Hussain, Asghar and Masood [5] review previous works which deals with the issues at the pre-processing stage and discuss proposed techniques to overcome those issues. Citing [11] that since the 80% of time spent on WUM is spent on pre-processing raw data, the authors **argued [expand:acknowledge]** that considerable attention should [expand: entertain: booster] be given to address problems at this stage in order to ensure accuracy of later phases of WUM. Both papers present [contract: proclaim: endorse] studies on issues specific to steps involved in pre-processing phase which are data cleaning, data filtering, user and session identification and path completion. The papers analyze and evaluate proposed techniques.

Pabarskaite & Raudys [11] **provide [expand:acknowledge]** a more extensive review of WUM papers prior to 2005 which covers literature related to both the pre-processing and pattern discovery phases. Most of the rest of the reviews investigate studies of WUM on specific applications. For example, Lappas [7] surveys **studies on[expand:acknowledge]** WUM application in the areas which have direct impact to society such as e-government, eeducation and e-politics. These are areas which were given less attention by WUM research community compared to business and computing. **Review on [expand:acknowledge]** the application of WUM to facilitate the prediction of future user request is done by Patil[6]. Accurate and efficient prediction of users future request will be able to overcome the propagation delay in heavily visited web sites. The author **analyzes [expand:acknowledge]** a number of proposed techniques and highlights related future research issues.. Ngai reviews studies of WUM in customer relationship management [12] and fraud detection[13].

Our review of WUM literature focuses on the learning algorithms applied for pattern discovery, giving additional emphasis on sequence discovery methods. Since pattern discovery is an essential phase of WUM, we believe [contract: proclaim:pronounce] that it is pertinent to give due attention to pattern discovery algorithms as the algorithms have a direct impact to the accuracy and quality of resulting patterns. Furthermore [expand: entertain: booster], since user web access transactions are temporal in nature and sequential pattern techniques yield more accurate results[14], in this review extra attention is given to studies involving learning algorithms which utilize sequential pattern methods.