Enhancement of “Technique for Order Preference by Similarity to Ideal Solution” Approach for Evaluating the Web Sources to Select as External Source for Web Warehousing

Hari Om Sharan Sinha, SC&SS, Jawaharlal Nehru University, New Delhi, India

ABSTRACT

The main concern of this paper is to evaluate the web sources, which are to be selected as external data sources for web warehousing. In order to identify the web sources, they are evaluated on the ground of their multiple features. For it, Multi Criteria Decision Making (MCDM) approach has been used. Here, among all the MCDM approach, the focus is on “Technique for Order Preference by Similarity to Ideal Solution” (TOPSIS) approach and proposing an enhancement in this method. The conventional TOPSIS approach uses Euclidean Distance to measure the similarity. Here, Jeffrey Divergence has been proposed to measure the similarity instead of Euclidean Distance which includes all the symmetric distances during computation. The Euclidean Distance only measures unidirectional distance whereas the Jeffrey Divergence includes multidirectional distances. Unidirectional distance includes only distance in one dimension but multidirectional distances includes differences, so more relevant in web sources evaluation. Experimental analysis for both the variations of TOPSIS approach have been conducted and the result shows the enhancement in the selection of web sources.

KEYWORDS

Enhanced TOPSIS, Jeffrey Divergence, MCDM Approach, TOPSIS Approach, Web Warehouse

INTRODUCTION

Web-warehouse consists of both the technologies viz. Web Technology and Data Warehouse Technology (Tan, Yen, & Fang, 2003; Ng et al., 1998). Comprehensively, Web-warehouse is an approach to develop the system, which have primary objective to identify, catalog, retrieve, store and analyze the data, available in the form of text, graphics, image, sounds, videos and other multimedia form, with the help of web technologies, in order to help the user to find and analyze the information effectively (Martinez et al., 2008; Tan et al., 2003). Web technology is basically an Internet technology. Nowadays modern society is the era of digital information (Singh et al., 2017a) where Internet is the prominent source of information. This information era leads to exponential and dynamic growth of the data on Internet, which provides voluminous data on Internet (https://www.crcpress.com/Mining-Multimedia-Documents/Karaa-ey/p/book/9781138031722) for individual, decision support and various research. So, most of the data is retrieved, transformed and stored from internet for inference, decision support system and data analytics.

DOI: 10.4018/IJNCR.2017010101

Copyright © 2017, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.
Despite being Internet a prominent source of information as well as an open platform to share and retrieve the data, the data available on web is not properly structured. So, it is not acceptable by traditional data warehouse. But data analytics requires more data for decision support system, so induces the traditional data warehouse to update itself to web warehouse (Zhu, & Buchmann, 2002; Ng et al., 1998; Singh et al., 2017b).

To design a web-warehouse, the architect has to tackle many challenges because of strict nature of warehouse (Immon, 2005; Ponniah, 2001) and open nature of Web. Web-data has dynamic and complex nature. Besides there are millions of web sources available on web. So, to find the relevant and coherent data on web is alike searching a needle in a stack of hay. Thus, the very first task of web-warehouse approach is to find out the relevant web sources as external data sources for warehousing. To ascertain the relevance of the web sources, they are assessed using numerous features. These features have been classified into three categories viz. web source stability, web data quality and contextual issues of web data (Zhu, & Buchmann, 2002). As we know, MCDM (Velasquez & Hester, 2013; Triantaphyllou et al., 1998) is an approach to find out the best among all the alternatives using multiple features. Before comprehensive explanation of MCDM, little more description of the set of features is here in this section.

The first category of features set explicates that along with numerous availability of web sources, the web data changes randomly as well as frequently and a large number of web sources are summing to the web. Thus, existing web source may change or disappear (Zhu, & Buchmann, 2002).

The second category of features set elaborates the quality of web data, since a large section of available data is not checked precisely before making it available on web. Since web is an independent and open platform. So, inconsistent, ill-structured, incomplete and wrong data is often available on web (Zhu, & Buchmann, 2002).

The third category of features set explains the context of data. The data available on web is browsing oriented not analytics oriented. Here Context of data explains not only challenges in the terms of relevance of data, but it also explains the easiness in extraction of data as well as metadata such as data definition, data derivation etc. (Zhu, & Buchmann, 2002).

All these features are taken into account to evaluate the web-sources, while designing an efficient web-warehouse. Zhu et al. proposed features, as evaluation parameters and used MCDM approach for evaluation of web-sources (Zhu, & Buchmann, 2002). In this paper, among all the MCDM approach, the main focus is on TOPSIS approach and especially on the enhancement of TOPSIS approach (Velasquez & Hester, 2013; Triantaphyllou et al., 1998; Zionts, 1979) by using the Jeffrey Divergence measure at the place of the Euclidean Distance measure (Ullah, 1996; Johnson & Sinanovic, 2001; Cha, 2007; Bayarri & García-Donato, 2008).

Remaining part of the paper is arranged as: Section 2 presents the explanation of the various features which are used for evaluation of web sources. Section 3 explains the MCDM approaches, here the main focus is on TOPSIS approach. The significance of Jeffrey Divergence method has been shown in Section 4. The proposed work has been explained in the section 5. In Section 6, the experimental system arrangement and result analysis have been shown. At last, the Section 7 has conclusion of the paper.

THE EVALUATION FEATURES

In this section, the aforementioned three categories of evaluation features have been refined further. The “Web Source Stability” has been categorized into Accessibility Durability, Availability and Refreshing rate. The “Web Data Quality” has been categorized into Origin, Objectivity, Accuracy, Completeness, and Metadata. Here Origin usually refers to data lineage. Lineage defines where the data come from, how it was derived and updated over time, so defines the data stream for web-warehousing (Ikeda & Widom, 2009). It also reflects the reliability of data and faith of one’s in it. Similarly, the Contextual Issues of web data has been classified into Relevancy, Timeliness and Layout (Zhu, & Buchmann, 2002).
The Planning Net: A Structure to Improve Planning Solvers with Petri Nets
www.igi-global.com/article/the-planning-net/126481?camid=4v1a

From Cochlear Implants and Neurotology to Brain Computer Interfaces: Exploring the World of Neuron Synapses for Hearing Impairments
Dionysios Politis, Miltiadis Tsaligopoulos and Georgios Kyriafinis (2016). *Handbook of Research on Natural Computing for Optimization Problems* (pp. 988-1015).
www.igi-global.com/chapter/from-cochlear-implants-and-neurotology-to-brain-computer-interfaces/153851?camid=4v1a