Assessing web sites quality: A systematic literature review by text and association rules mining

CrossMark

Rim Rekik[a,*], Ilhem Kallel[a,c], Jorge Casillas[b], Adel M. Alimi[a]

[a] REGIM-Lab: Research Groups in Intelligent Machines, University of Sfax, ENIS, BP 1173, Sfax, 3038, Tunisia
[b] Department of Computer Science and Artificial Intelligence, University of Granada and the Research Center on Information and Communications Technology (CITIC-UGR), Granada, E-18071, Spain
[c] ISIMS: Higher Institute of Computer Science and Multimedia of Sfax, University of Sfax, Tunisia

ARTICLE INFO

ABSTRACT

Nowadays society is deeply affected by web content. A web site, regardless of its category, can provide or not for users their needs. To identify its strengths and weaknesses, a process of analyzing and assessing its quality, via some criteria, is necessary. Assessing web sites is considered as a Multiple Criteria Decision Making problem (MCDM), with a massive number of criteria; a reduction phase is needed. This paper presents, firstly a Systematic Literature Review (SLR) to identify the purposes of recent researches from the assessment and determine the affected categories; secondly, it proposes a process of collecting and extracting data (criteria featuring web sites) from a list of studies. Text mining is applied for this SLR to construct a dataset. Then, a method based on Apriori algorithm is assigned and implemented to find association rules between criteria and the category of the web site, and to get a set of frequent criteria. This paper also presents a review on soft computing assessing methods. It aims to help the research community to have a scope in existing research and to derive future developments. The obtained results motivate us to further probe datasets and association rule mining.

1. Introduction

Over the last decade, the circulating online data in the World Wide Web has been rapidly increasing in terms of amount and diversity. In addition, most of the internet visitors used those data without knowing their credibility. In order to guarantee the reliability and credibility of online data, several assessment solutions have emerged.

In the field of web assessment, evaluating the quality of web sites is a key to filtering interesting sites for surfing or not. A web site is regarded not only as a showcase for commercial goals but also as a scientific showcase for considering an institution or represented organization's reliability. Assessing fields is complex, due to the panoply of web sites such as E-commerce, education, entertainment, health, etc. and also the diversity of criteria qualifying a web site. In some cases, the most common strategy adopted by some approaches is to know the type of web site to propose an interesting set of criteria in the field of assessment. For example, a commercial web site that provides payments mode (Chen, Rungruengsamrit, Rajkumar, & Yen, 2013) will not be evaluated with the same factors as an educational one (Violante and Vezzetti, 2015). In other cases, some studies (Ozmen-Ertekin and Ozbay, 2012) proposed a general model to assess any type of web site

by choosing common characteristics. Consequently, the problem is one of multiple criteria decision making (MCDM). Existing methods are mainly based on making a hierarchy to divide high level criteria, sub-level criteria and alternatives. In fact, each criterion is weighed but there is no assumption about any association between them.

1.1. Research purpose and motivation

One such topic of web sites' quality assessment is a critical problem in practice and research. Therefore, the first motivation is to identify the current state of the art by applying a systematic literature review (SLR) as introduced by Kitchenham et al., (2010) and Kitchenham and Charters (2007), and recently applied (Aleti, Buhnova, Grunske, Koziolek, & Meedeniya, 2013; Yagüe et al., 2014). Its aim is to find the goals of assessment in this field and to classify studies according to web sites' category in order to determine the different and important domains concerned by the evaluation. In general, it gives a good overview for exploring the existing research and expanding new horizons in the area. Moreover, text mining for SLR induces us to collect data from studies and apply association rules mining (see Appendix A for more details about this concept). The purpose is to search for criteria

---

* Corresponding author.
  E-mail addresses: rim.rekik@ieee.org (R. Rekik), ilhem.kallel@ieee.org (I. Kallel), casillas@decsai.ugr.es (J. Casillas), adel.alimi@ieee.org (A.M. Alimi).
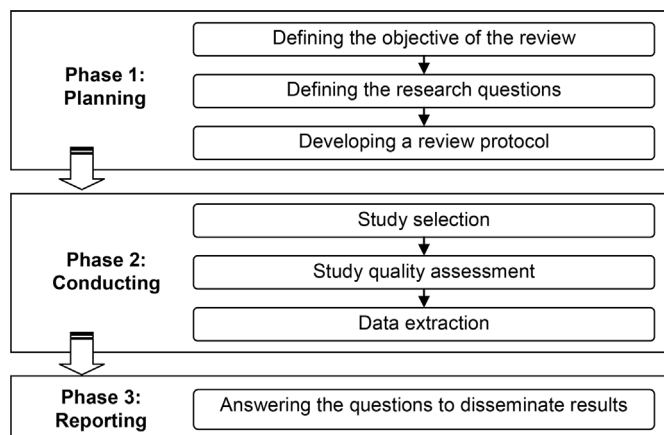
Fig. 1. Summary of the SLR process phases according to (Kitchenham and Charters, 2007).

dependencies with the category of the web site and to know the most reliable ones. Moreover, another motivation is to search for approaches that use soft computing techniques to resolve the problem.

1.2. Research method

In this work, we exercise the SLR approach according to Kitchenham and Charters' guidelines (Kitchenham and Charters, 2007) supervised by Fig. 1. The process is composed of three phases: planning, conducting and reporting the review. The aim of the first phase "planning the review" is to define the objective of the SLR and a clear review protocol. It specifies the main raised research questions, the adopted search strategy and a set of established inclusion and exclusion criteria to select a publication. The second phase "conducting the review" is for executing the protocol. In the third phase, we report the obtained results.

1.3. Content overview

The remaining part of the paper is organized as follows: Section 2 introduces a detailed description of the systematic literature review protocol. Section 3 presents conducting the review when applying the developed protocol. In Section 4, a report of the extracted data and related works from the previous review is discussed. Finally, some findings are drawn in Section 5 for developing future work and the main conclusions are given in Section 6.

2. Planning the review

In planning the review phase, the objective of the SLR and the research questions are defined. Moreover, a clear review protocol is developed. It consists of defining a search process strategy and the inclusion/exclusion criteria considered in the research.

Fig. 3. Example of a Lucene search expression.

2.1. Objective and research questions

The objective of the SLR consists of exploring different aspects of the assessment of web sites quality existing in the literature. According to the objective and motivation of the review as described in section 1.1, we formulate a set of six research questions as follows:

RQ1: What are the purposes of the recent research from the assessment of web sites quality?

RQ2: What are the most common categories of web sites considered for the assessment?

RQ3: What are the criteria that characterize a web site? What are their semantic groups?

RQ4: Can we extract association rules between the criteria? Which ones?

RQ5: What are the frequent criteria considered in an assessment process?

RQ6: Due to the subjectivity and imprecision of that MCDM problem, are there studies that performed the assessment using soft computing or hybrid methods? If so, in which phase of the evaluation and what is the motivation of applying such intelligent methods?

2.2. Search process

It is important to follow a search strategy in order to ensure a convincing review conducted in phase 2 (see Fig. 1). In fact, a phase is needed for exploring scientific publications from related journals and conferences in relevant electronic sources such as Elsevier's Scopus, Elsevier's ScienceDirect, IEEExplore, ACM Digital Library, SpringerLink or Google Scholar. It is necessary to define some key concepts as selection words. Indeed, we consider some words such as "assessment", "assessing", "evaluation", "quality", "web sites", etc. A combination of these terms should be made to enlarge the scope of searching for better results.

2.3. Inclusion and exclusion criteria

Since we cannot include all collected papers, we introduce some inclusion and exclusion criteria. In order to select the most relevant ones, we determine the criteria that specify whether a study will be included or excluded. The first inclusion criterion based on terms which appeared in the titles, abstracts and keywords in studies by browsing the computer science discipline; an identification of relevant ones was established. However, papers published before 2009 and non-English written studies have to be excluded. In addition, we exclude some sub-disciplines such as "Web services" not related exactly to the topic of assessment. Fig. 2 presents a boolean expression query performed using the Scopus database.

After obtaining a large set of papers, a step to eliminate short ones (up to 4 pages) and to make a fine filter on the set of papers is required

Fig. 2. The performed search query in Scopus.