Web Data Conceptual Framework: Integration, Cleaning, Analysis, Visualization, and Security

Fethia Benhadj Djilali Magraoua and Saliha Hafifi

Abstract The rapid growth of the web in the last decade makes it the largest publicly accessible data source in the world. The amount of data/information on the web is huge and still growing. The web has many unique characteristics, making mining helpful information and knowledge fascinating and challenging. The coverage of the information is also extensive and diverse.

Keywords Web data · Integration · Cleaning · Analysis · Visualization · Security

1 Introduction

With millions of customers now online, the importance of websites in influencing their purchasing decisions is significant. With the company's website having the potential to ideally become a single all-encompassing access point to all the stakeholders – customers, investors, employees, and external partners, the management of their perceptions and the website has become important for business success. The unique characteristics of the Internet, such as intense competition, immediate access to product and service information, instant price comparisons, and the ease with which customers can leave an e-commerce website, force companies to concentrate on the management and measurement of this critical customer interface. Knowledge discovery is a term used in databases to describe the process of analyzing data (KDD). Discovery of useful patterns or knowledge from data sources is a common definition. Data mining is a multidisciplinary field involving machine learning, statistics, databases, artificial intelligence, information retrieval, and visualization. Numerous data-mining projects can be found in the market today. For example, supervised learning (also known as classification), unsupervised learning (also

F. Benhadj Djilali Magraoua (✉) · S. Hafifi
University of Djilali Bounaama Khamis Miliana, Khemis Miliana, Algeria
e-mail: f.ben-magraoua@univ-dbkm.dz; s.hafifi@univ-dbkm.dz

S. Sedkaoui et al. (eds.), *International Conference on Managing Business Through Web Analytics*, https://doi.org/10.1007/978-3-031-06971-0_1

known as clustering), association rule mining, and sequential pattern mining are among the most common. The web data principles are not a specific set of technologies, but rather simple principles and patterns.

2 Ethical Issues in the Analysis of Web Data

As a result, data mining raises significant ethical concerns because individuals who aren't made aware of the collection and use of their personal information aren't given the option to give their consent or withhold it. On the Internet, it's common practice to gather data invisibly. When personal data is misused or used for a purpose other than the one for which it was provided, mining the web can put people at risk (secondary use). This knowledge, on the other hand, has numerous benefits. Planning and control applications benefit greatly from the knowledge gleaned through data mining. Web data mining has a number of specific advantages, such as enhancing the intelligence of search engines. Analyzing a web user's online behavior and turning it into marketing knowledge are other ways to use web data mining in the context of marketing intelligence.

There may be ethical concerns when mining web data that does not include personal information, such as data on automobiles or animals. This chapter, on the other hand, focuses on web data mining that includes some personal data. Only harm to individuals will be examined; any harm to organizations, animals, or other subjects is outside the scope of this investigation. For now, we'll focus on the private sector's web data mining applications. Because of this, personal data mining on the web will be analyzed from an ethical standpoint. This technique has a lot going for it, and we think it has a lot of good qualities and potential. Web data mining is appealing to businesses for a variety of reasons. Consumer data and government records, for example, could be used to determine who might be a new customer and other useful information. In the most general sense, it can increase profits by actually selling more products or services or minimizing costs. Marketing intelligence is needed to accomplish this. It's possible to use this intelligence to better your marketing strategies, competitor analyses, or interactions with clients and customers. The different kinds of web data related to customers will then be categorized and clustered to build detailed customer profiles. This helps companies retain current customers by providing more personalized services and contributes to the search for potential customers. That web data mining can be very beneficial to businesses is beyond dispute. To ensure that this technique will be further developed in a properly thought-out way, however, we shall focus on its possible objections. For a well-informed development and well-considered application, being aware of all the potential dangers is critical. The different ways in which privacy is threatened are the dangers of web data mining. To structurally analyze the many different ways to mine the web, it's important to distinguish between the various types of web data mining. Web structure data, such as hyperlink structure, can be distinguished from actual data on web pages and across web documents and weblog data regarding the users who browsed web pages. We shall divide web data mining into three categories:

1. Analyzing content data found in web documents falls under the heading of "content mining." This can be anything from a picture to an audio file. Content mining, on the other hand, will only refer to text mining in this study.
2. Structure mining is a subcategory of link mining. Its goal is to examine the relationships between various web documents.
3. In the field of data mining, this is known as "usage mining." Data logged by users when they interact with the web is analyzed by usage mining.

"Log mining" is a term used to describe the process of extracting data from web server logs. When used in conjunction with some form of content mining to decipher the contents of hyperlinks, structure mining can be even more valuable.

3 Security, Privacy, Access Control, and Sharing

Unlike other data mining techniques, web usage mining has a unique application. We'll talk about it separately because it has a different set of advantages and challenges values in a different way.

3.1 Privacy Threatened by Web Data Mining

The use of personal data in web data mining can disrupt some important normative values. This is what we are going to discuss in this section. People's (informational) privacy may be violated, which is a clear ethical concern. Protecting the privacy of users of the Internet is an important issue. Privacy, on the other hand, is conceptually weak. In today's society, the term "privacy" covers a wide range of social practices and concerns. The philosophical and legal debates surrounding privacy will be omitted from this chapter. We'll use a looser (and more common) definition of informational privacy for the purposes of this discussion. In order to maintain one's privacy online, one must be able to manage the information that is made public about oneself. Information about an individual's privacy is safeguarded using this term. When information about an individual is obtained, used, or disseminated without their knowledge or consent, their privacy may be violated. Web mining privacy concerns frequently fall into this category. As a result, we'll be focusing our attention on this section. It is in this context that the term "privacy" will be used throughout the rest of this chapter. However, the value of "individualism" may be violated if people are judged and treated based on patterns found through web data mining. To begin, let's examine the connection between privacy and individualism more closely. The privacy of an individual may be violated when data is gleaned from the web through web data mining. People's privacy may be violated if their data is categorized and grouped into profiles before being used for decision-making. In this case, however, the discovered information is no longer linked to specific individuals, and no direct sense of privacy violation because the profiles do not contain “real” personal data is violated when the data is anonymized before being produced. Group profiles, on the other hand, can be used as if they were personal data, resulting in the unfair evaluation of individuals – known as individualization (see the following section). Privacy can be thought of as a stepping-stone to other fundamental values. As Vedder (2000: 452) puts it, “... privacy is a servant of many master values.” Categorical privacy, which would allow group characteristics that are applied as if they were individual characteristics to be considered personal data, could be a solution. Such a solution, according to Tavani, is not appropriate because it may necessitate the creation of new privacy categories as new technologies are introduced.

3.2 Individuality

One way to describe the quality of individuality is to say that it is the quality of being an individual or of having a distinct personality from others. Individualism is a strong Western value. The core values of being an individual and expressing one’s individuality are widely held in Western countries. A tendency to judge and treat people on the basis of group characteristics rather than their own individual characteristics and merits can result from profiling through web data mining. 10 A person’s sense of self is jeopardized if group profiles are used as a basis for policymaking or if they are made public in some other way. As a result, individuals will be treated less like individuals and more like members of a group. The risk is heightened when profiles contain personal information that should be kept private and are, for example, used in allocation procedures to make decisions. People may be stigmatized or discriminated against simply because they are members of a group or because they have certain characteristics. The use of factors like race and religion in making decisions can be both inappropriate and discriminatory. Non-distributive group profiles pose an even greater threat because not every member of the group shares every characteristic of the group. Using probabilities, averages, and other statistical concepts, non-distributive group profiles often obscure personal information. It is no longer possible to identify an individual from the anonymized information because it no longer contains any data that can be used to identify them.

4 Web Analytics Features, Benefits, and Limitations

All of these advantages demonstrate that web data mining is an extremely valuable technique that is being developed and applied on a large and growing scale. However, there are some serious threats to some of the most important values in the web data mining field, and this is likely to cause a lot of tension. Unfortunately, many business professionals who use web data mining do not see any ethical issues with it. Twenty web data miners were interviewed in order to get a better understanding of current practices and the attitudes of web data miners toward ethical issues. Using interviews and a literature review, we can conclude that people prefer to discuss the benefits of web data mining rather than the possible risks. According to them, web data mining does not pose a real threat to privacy and other values because of a variety of reasons. Arguments in favor of data mining's near-limitless use can be broken down into six categories, each of which contains valuable information. The purpose of this brief discussion is to demonstrate that the arguments presented here do not support the use of data mining indefinitely.

4.1 Limitation

Web data mining itself does not raise any new ethical issues for discussion or investigation.

Laws and online privacy statements guarantee the confidentiality of personal information.

Because so many people have opted to give up their privacy, why not make use of it?

Most of the information gathered is of a non-personal nature or is used to create anonymous profiles.

There are fewer unsolicited marketing approaches as a result of web data mining.

Personalization leads to individualization instead of de-individualization.

5 Information Diffusion on the Web

With the advent of web, social networks have become an important medium for the dissemination of information in the Internet. The process of information dissemination in social networks has been studied using a number of information diffusion models. Due to the collaborative nature of the networks and limited accountability of the users, the media is often misused for spread of rumors and misinformation. In this chapter, we have a proposed a novel information diffusion model for the spread of misinformation using evolutionary game theory and evolutionary graph theory. The proposed model could be used to analyze as well as predict the spread of misinformation. It also provides a framework to study the effects of multiple campaigns in the network which would enable us to estimate the efficacy of launching countercampaigns against the spread of misinformation. We have used extensive simulation to support our claim.

Sharing information, such as news and rumors, is a key feature of social networks. Using natural connections, information can be disseminated in written, oral, or electronic form. With the widespread adoption of the Internet and the World Wide Web, the physics of information diffusion has changed. It used to be difficult for people to spread information in a large community because of the high costs associated with deploying the necessary technology to reach a large number of people. This stumbling block has been largely dismantled, thanks to the widespread availability of high-speed Internet. Due to its importance in social interactions and day-to-day life, information diffusion has been one of the primary research topics in the field of social network studies. It has only been in the last 20 to 30 years or so that there has been a shift toward actively participating in and shaping the flow of information and innovation. We can reason about the spread of information by modeling the diffusion of information in networks.

6 Important Approaches to Web Measurement

Businesses now view websites as more than just another channel or storefront or a simple informational portal for their customers. If you don't have an effective website, you're missing out on a lot of business opportunities. It is much easier for businesses to make adjustments and enhance their operations if they can get early and frequent feedback on how their website is performing from the point of view of its users. Several instruments and methodologies were developed to measure the website performance, usability, and quality in information systems, marketing, and operations management literature. This study reviews the literature in web quality measurement and employs a 25-item instrument developed by Aladwani and Palvia to measure the user-perceived web quality. It attempts to test the factorial validity of the instrument in Australian context using structural equation modeling technique. Analysis revealed that the data set do not fit the Aladwani and Palvia's model well enough.

6.1 Background and Literature Review

Many businesses are known only through their websites on the Internet. Whatever the size of a company, whether it's a sales brochure, or a customer contact point, or the sole distribution channel for the product, creating and maintaining an effective website is essential in today's business world. Though in the early days of Internet commerce, websites were expected to provide some entertainment to the customers, it is now considered irrelevant in today's business environment, except in some entertainment service websites. E-commerce websites' ultimate goal is to draw in potential customers and encourage them to make a purchase. In order to have a successful website, it must reflect the company's value proposition and meet the needs of its customers. If you think about it, the business strategy and operational policies of any given company are reflected in their website. Attracting and converting visitors into customers begins with the quality of the website and the way it interacts with its users. Despite the fact that companies have spent a lot of money advertising their websites, only 3.5% of the unique visitors buy something. Customer loyalty and recurring revenue can be increased by providing a superior online experience. Order fulfillment is a top priority for customers who use the Internet to gather information or make a purchase. For companies that conduct business online, additional complexity is required in terms of security, backup, and redundancy. Quality support for various functions, such as information search, transactions for purchasing goods and services, and post-sale support, is critical to the effectiveness and overall quality of a website. Operations management literature's concept of quality as "fitness for use" and the role of users or consumers in determining it are adopted in information systems research as well.

6.2 Validity and Correlations

Aladwani and Palvia's four-dimensional construct of perceived web quality and its validity are examined by looking at the relationships between scale ratings and Amazon users' overall quality ratings (Table 1).

It can be seen from the table above that the correlations between all four factors range from 0.526 to 0.657 and are all significant. However, these four factors/constructs have a significant correlation with the Amazon website's overall quality rating (global quality), which ranges from 0.298 to 0.574. Users' perceived web quality (the sum of all 25 items' scores) has a strong correlation with each of the 4 factors, with coefficients ranging from 0.777 to 0.869. Additionally, there is a statistically significant correlation between the overall quality rating of the website and the perceived web quality index. As a result, the accuracy and reliability of the test are further supported by this research.

Table 1 Correlations among factors and statistics

Factors	Technical adequacy	Content quality	Specific content	Appearance	User-perceived web quality	Overall quality of Amazon
Standard deviation	0.76	0.916	0.98	0.96	0.76	1.02
Content quality	657					
Specific content	.601	621				
Appearance	619	.526	.526			
User-perceived web quality	.869	.823	.799	.777		
Overall quality of Amazon	.574	493	.412	.298	.517	
Mean	5.02	4.85	5.03	4.62	4.83	4.94

7 Conclusion

The web was a hot topic in some research fields, but it had yet to be popularized and become a common technology. Some areas of the Semantic Web are already taking advantage of its potential, such as search engines and the metadata embedded in the pages of websites that want to be better understood by robots. The Semantic Web is currently in an intermediate stage. The Semantic Web is still in its infancy, and the development of applications that can take advantage of this new model is just beginning. Massive data publishing is still going on, but only a small number of applications are taking advantage of it. The importance of applications, or apps, as they are known today, is undeniable in the world of mobile devices. It's no surprise that many people in big cities use published data mined by an app when they're looking for a cab.

References

Amazon: Amazon.com launches web services [Press Release] (2002, July 16). Retrieved from http://phx.corporate-ir.net/phoenix.zhtml?c=176060&p=irol-newsArticle&ID=503034. Google Scholar

Rice, M.: What makes users revisit a web site. Market. News. **31**(6), 23 (1997)