International Journal of Scientific & Engineering Research, Volume 3, Issue 12, December-2012
ISSN 2229-5518

Knowledge Discovery on Users' Requirements for Building B2C Websites Using Data Mining Techniques

Houzifa M. Salahdeen Hintaya [a], Faudziah Ahmad [b]

Abstract— The purpose of the study was to identify users' requirements for developing a B2C website using data mining techniques. A 30-item questionnaire for evaluating users' satisfaction of a web site was developed. Descriptive analysis and data mining techniques were used for analysis. The criteria investigated were clarity, objectivity, content, currency, responsibility, navigation, author, accuracy, design and stability, coverage, accessibility, purpose, and reliability. A data mining approach consisted of four stages was proposed.

Index Terms— *data mining, reduct, website development criteria, website evaluation.*

1 INTRODUCTION

Electronic commerce specifically B2C (Business to Consumer) has been a popular medium for users to obtain information about products and services, and/or conduct buying and selling transactions [1][2]. Thus, having a quality B2C website is crucial in attracting online shoppers to visit a company's online store and learn about its products and services and in ensuring repeat purchases.

Despite the increased use of the Internet as a shopping channel, many researchers, however, have reported that the number of online shoppers and total sales through the Internet are still marginal as compared with those in traditional retailing. For example, there are many online bookstores available on the Internet; however, not many users are using the online facilities to purchase books. This might be due to reasons such as poor website design, slow loadings of web pages, unattractive interfaces, or other related matters hindering users to adopt the electronic technology for purchasing purposes [3][4][5].

According to [3] website facilities do significantly influence attitudes towards online purchasing intention as users choose to visit a website if the websites fulfils their criteria. Some good criteria that are used in evaluating websites are clarity, accessibility, coverage and reliability. Since Internet users differ from one another in their selections, this study thus, attempts to identify users' requirements for developing a bookstore website.

The data mining has been known as a strategic a tool that can used to improve businesses [6]. For example, data mining techniques can be used to understand interactions and relationships data elements in a dataset. Data mining (DM) has been known as a technique to support the process of redesigning a business by extracting knowledge hidden in large volumes of data [7]. Business intelligence (BI) tools such as data warehousing, data mining, and OLAP can be used to discover knowledge on customer behaviours that can be used by retailers to develop plans based on users' needs and desires identified.

In this study, data mining techniques were used to explore the possibility of selecting important criteria for developing a bookstore website. This study evaluated five bookstore websites and then identified users' preferences for a bookstore websites. It was intended to help developers identify a good set of requirements for designing an online bookstore website. Specifically, the objectives were to identify users' requirements for online bookstore websites development, rank the importance of the criteria and present a data mining approach for identifying the important criteria.

- *Houzifa M.Salahdeen Hintaya is currently a postgraduate student at the School of Computing, UUM College of Arts and Sciences, Universiti Utara Malaysia, Kedah, Malaysia. (e-mail: s809389@student.uum.edu.my)
- Faudziah Ahmad is a lecturer at the School of Computing, UUM College of Arts and Sciences, Universiti Utara Malaysia, Kedah, Malaysia (Tel: 6049284787; fax 6049284753; email fudz@uum.edu.my).

2 RESEARCH METHODOLOGY

2.1 Websites

Five bookstore websites have been evaluated. The website links were http://www.mph.com.my/, http://www.pelangibooks.com/., http://www.rusabooks.com/, http://www.silverfishbooks.com/, and http://www.bookcafe.com.my/.

2.2 Data collection and preprocessing

Questionnaires were used to gather data. The questionnaire was constructed based on 13 criteria identified from past literatures. These were clarity, objectivity, navigation, content, currency, author, design and stability, accuracy, responsibility, accessibility, coverage, purpose, and reliability.

Table 1 shows description of each criteria.

TABLE 2. CRITERIA AND DESCRIPTION

No	Criteria	Description
1	Clarity	The web visitor can understand the information presented on the website
2	Objectivity	The website was built to achieve certain objectives/ goals
3	Navigation	The website was helpful in guiding visitors to move from one page to another
4	Content	The website contained enough information such that visitors can make decisions
5	Currency	The website contained up-to-date information
6	Author	The website provided information such as professional qualification on the website's developer
7	Design and Stability	The website was simple, functional and easy to use
8	Accuracy	The website provided correct information
9	Responsibility	The website published information that are legal and follow ethical standards
10	Accessibility	The website allowed users to visit at any time and without any problems
11	Coverage	The website provided comprehensive information
12	Purpose	The purpose of the information presented on the website was clear
13	Reliability	The website published information that were obtained from reliable sources

Respondents consisted of academic staff and students (undergraduate and postgraduate) at University Utara Malaysia (UUM). A total of 110 questionnaires were collected. 50% of the returned questionnaires were from the undergraduate students, 30% were from the postgraduate students and 20% were from UUM staff. Data gathered were keyed into a database for preprocessing.

Data preprocessing involved cleaning the data i.e ensuring that the data is free from missing values, noise (contain errors, outlier values) and inconsistencies (discrepancies of units used). In this study, missing values were replaced with mean value because this method has been commonly used by many researchers [8]. After preprocessing, a complete dataset was obtained and used for experiments.

3.3 Experimental process

The process involved several steps (Figure 1).

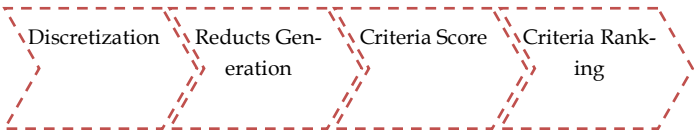

Figure 1. Process of identifying important criteria

Stage 1 Discretization

The dataset was divided into two parts: training and testing. The training set was Frequencies were used for calculating the occurrences of variables in reducts. GA, RSES, 0.20, and 2, best techniques for reduction, classification, split factor and ring seed were used to generate reducts.

Stage 2 Generate reducts

Several experiments were conducted to identify suitable data mining techniques for reduction, classification, split factor and ring seed. The techniques tested for reduction were Genetic Algorithm (GA), Johnson (JA) and Holte's 1R (1R). GA is a search technique used in computer science to find approximate solutions to optimization and search problems [9]. JA is a way to solve the all-pairs shortest path problem in a sparse, weighted, directed graph [10], while, 1R returns all singleton attribute sets [11]. Output from the three techniques were sets of attributes (known as reducts) that were found to be relevant or important to the dataset. The accuracy of a reduct was measured in terms of percentage of accuracy. Standard Voting (SV), Voting with object tracking (V), and Standard/tuned voting (RSES) algorithms were tested for best classification technique. Other parameters analyzed were splitting factor (percentages of data for training and testing) and random seed number ranging from 1 to 10. The experiments conducted showed that GA, RSES, 0.20, and 2 were best techniques for reduction, classification, split factor and random seed. The experimental results were not shown in this paper due to space limitation.

- Stage 3 –Criteria Score

This step analyzed 110 gathered samples. SPSS version 13.0 has been used for descriptive statistics. Total frequencies were used for calculating scores of variables in reducts. GA, RSES, 0.20, and 2, best techniques for reduction, classification, split factor and ring seed were used to generate reducts.

- Stage 4 Criteria Ranking

The criteria were ranked based on the scores identified in Stage 3. Scores of of 50% and more were considered as important and were ranked from the most important to the least important order.

3. FINDINGS AND DISCUSSION

This section presents the set of criteria identified for an online bookstore websites, the set of criteria ranked based on its importance, and proposed data mining approach for identifying the set of criteria.

Table 2 shows the results. In the table, Occur denotes the number of times the variable occurs in all reducts (total is 183 reducts), T.Occur is the sum of Occur of variables for the criteria, Rnk is the ranking position of the criteria. Rnk of 1 denotes the most important, 2 represents second most important, while, 13 is the least important.

TABLE 2. CRITERIA AND PERCENTAGE OF OCCURRENCES

NO	Criteria	Variables	Occur	T.Occur	Rnk

1	Clarity	1a	90		
		1b	10		
		1c	18		
		1d	21		
		1e	13		
		1f	52	204	1
2	Objectivity	2a	40		
		2b	42	82	2
3	Navigation	3a	9		
		3b	38		
		3c	0	47	6
4	Content	4a	34		
		4b	12		
		4c	10		
		4d	4	60	3
5	Currency	5a	5		
		5b	18		
		5c	24		
		5d	11	58	4
6	Author	6a	8		
		6b	9		
		6c	22	39	7
7	Design and Stability	7a	6		
		7b	1		
		7c	21		
		7d	9	31	9
8	Accuracy	8a	3		
		8b	14		
		8c	18	35	8
9	Responsibility	9a	8		
		9b	12		
		9c	15		
		9d	16	51	5
10	Accessibility	10a	15		
		10b	14	29	11
11	Coverage	11a	8		
		11b	8		
		11c	9		
		11d	5	30	10
12	Purpose	12a	2		
		12b	6		
		12c	7		
		12d	8		
		12e	7	23	12
13	Reliability	13a	6		
		13b	5		
		13c	3	14	13

Based on Table 1, it can be seen that clarity is the most important criteria for online bookstore websites. The second important criterion is objectivity while the least important is reliability.

3.3 The proposed data mining approach for identifying the set of criteria

Fig. 1 shows the approach. The approach consisted of four stages: Stage 1) instrument development and data collection; Stage 2) preprocessing and selection of suitable data mining techniques; Stage 3) criteria scores; and Stage 4) criteria ranking.

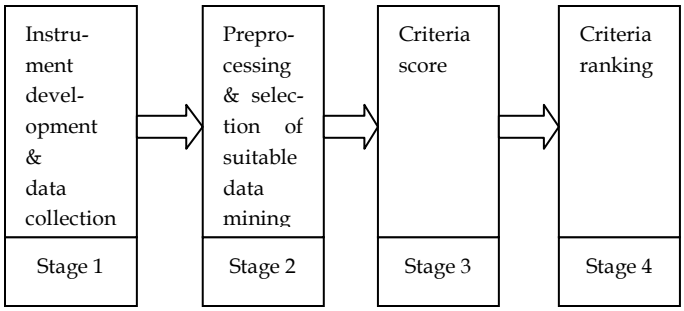

Fig. 2. Proposed data mining approach

The approach has been tested and was able to select and rank important criteria.

4. Conclusion and Future Work

The study has successfully achieved all objectives. The important criteria based on its ranking from most important to least important were clarity, objectivity, content, currency, responsibility, navigation, author, accuracy, design and stability, coverage, accessibility, purpose, and reliability. The proposed data mining approach used for the selecting the criteria consisted of four stages: instrument development and data collection; preprocessing and selection of suitable data mining techniques; important criteria identification; and criteria ranking. Suggestions for future works include exploring other different techniques to select important criteria, and comparing accuracy of different techniques used.

Acknowledgements

The authors would like to thank all colleagues and students especially Khaled Ali Othman Ali who contributed to this study.

References

[1] Korner V. & Zimmermann H. Management of customer relationships in business media (MCR-BM). *Electronic Markets* 2000, 10:162–168.

[2] Geissler GL. Building customer relationships online: the web site designers' perspective.*Journal of ConsumerMarketing* 2001, 18:488–502.

[3] Hoffman DL., Novak TP. & Chatterjee P. Commercial scenarios for the Web: opportunities and challenges. *Journal of Computer-Mediated Communication* 1995; 1:23–45. [WWW document]. URL http://www.ascuse.org/jcmc/vol1/issue3/hoffman.html.

[4] Jarvenpaa SL. & Todd PA. Consumer reactions to electronic shopping on the World Wide Web. *International Journal of Electronic Commerce* 1997;1:59–88.

[5] Lohse GL. & Spiller P. Electronic shopping: the effect of customer interfaces on traffic and sales.*Communications of the ACM* 1998;41:81–87.

[6] Kantardzic M. Data Mining: Concepts, Models, Methods, and Algorithms. Department of Computer Engineering and Computer Science (CECS), University of Louisville: John Wiley & Sons; 2003.

[7] Folorunso, O & Ogunde, AO. Data Mining as a Technique for Knowledge Management in Business Process Redesign. The Electronic Journal of Knowledge Management 2004; 2(1):33-44. [WWW document]. URL www.ejkm.com.

[8] Roderick JA, Little D, Rubin B. Statistical analysis with missing data, New York:John Wiley & Sons Inc; 1986.

[9] Whitley D. *A genetic algorithm tutorial*. Statistics and Computing; 4:65–85.

[10] Han J and Kramer, M. Data Mining: Concepts and Techniques, San Francisco: Morgan Kaufmann Publishers; 2001.

[11] Holte RC. Very Simple Classification Rules Perform Well on Most Commonly Used Datasets, *Machine Learning*, 1993; 11:63-92.