Evaluating E-Commerce Sites with AI Tools: a modelling approach

Antonia Stefani[1], Dimitris Kalles[2], Michalis Xenos[1],[2]

[1]School of Sciences and Technology, Hellenic Open University
[2]Laboratory of Educational Material and Educational Methodology, Hellenic Open University
{stefani,kalles,xenos)@eap.gr

Abstract

This paper proposes a method for modelling the quality aspects of e-commerce systems based on Bayesian Networks. The proposed model can be used as an instrument for the measurement of the quality of such systems. The paper presents the theoretical background of the model, as well as practical issues arising from its application. Special emphasis is placed on the model's structure and usage, as well as on the potential to extend it and validate it through further AI techniques.

Keywords: Evaluation, E-Commerce, Artificial Intelligence.

1. Introduction

E-commerce is now a mature field. This fact is confirmed by the increasing number of enterprises that invest into the creation of e-commerce systems, and the continuous expansion of economic and commercial transactions through the Internet. E-commerce can be defined as follows [Zwass (1996)]: sharing business information, maintaining business relationships, and conducting business transactions by the means of telecommunication networks. Depending on the type of transactions performed electronically, we usually differentiate between Business to Consumer (B2C) and Business-to-Business (B2B) systems.

In e-commerce systems, interaction with the end-user is conducted through web-based applications including a human-computer interface. Since all user-system communication is based on such interface, it is self evident that the quality of an e-commerce system is directly related to the quality of the human-computer interface through which the end-user interacts with the web-based applications. Usually, end-users value e-commerce systems that are flexible, usable, easily adaptable to their needs, and that offer a full range of applications. But how can one measure the quality of e-commerce systems and define the extent to which they meet end-users' requirements? To this end, it is necessary to define what constitutes a high-quality e-commerce system as well as a methodology for evaluating the quality of e-commerce systems [Burrows. (1999)]. This paper proposes a model based on Bayesian Networks that can be used as an instrument for measuring the quality of e-commerce systems, as well as defining specific quality requirements during the design process of e-commerce systems.

The rest of this paper is structured as follows. The next two sections present some background on quality evaluation and on Bayesian networks. Section 4 then describes the fundamental aspects of the model's analysis as they pertain to a BN problem formulation and Section 5 describes the implementation of the model. In Section 6 we cast the remaining aspects (and research directions) of model development and validation as applications of fundamental machine learning techniques. In Section 7 we summarize our contribution and briefly relate the emerging mainstream presence of several AI techniques to the recent abundance of generic AI tools.

2. Quality aspects background

Most e-commerce systems seek to provide high quality services to the clients. To this end they include specific applications to meet specific end-user requirements. Examples of such requirements are searching capabilities, flexible navigation or the ability to group goods and applications like search engines and navigation maps have been developed in order to meet such requirements. Even if the type of applications that an e-commerce system integrates changes in the future, the user requirements relating to the e-commerce system will remain unchanged. It is thus reasonable to conclude that the quality and evaluation methods of e-commerce systems will always be dependant on the quality of similar applications and their ability to meet end-user requirements. Such quality characteristics should be taken under serious consideration during the development of e-commerce systems.

Past approaches about the quality of e-commerce systems are emphasizing usability standards [Ivory and Hearst (2003)][Becker and Berkemeyer (2002)] using techniques like feature inspection methods or collecting data about end-users' opinion with questionnaires. These methods provide an important feedback to the researcher and their results can be utilized as a useful background for future work, however, they do not solidly contribute to a dynamic model. Other approaches evaluating web systems and e-commerce systems focus on statistical analysis of usage patterns [Mich et al. (2003)][Chang and Hon (2002)]. The importance of the proposed model lies in its dynamic character, as regards the formation of quality aspects' groups, as well as the potential to retroactively fine tune the model based on its application.

Note that the importance of each of the above mentioned quality characteristics depends on each e-commerce system's specificities as well as the end-user requirements and developer priorities for the specific system. For the scope of this paper, the development of the proposed model was mainly based on Business to Consumer (B2C) systems.

3. Bayesian networks background

The proposed model is based on the notation and formation of Bayesian Networks (also referred to as Belief Networks and, hereafter, referred to as BN), a special category of graphic models where nodes represent variables and the directed edges represent relations between nodes [Pearl (1986)], [Lauritzen and Spiegelhalter (1990)]. The fundamental working assumption about BNs is that we might be uncertain about the precise strength of the influences between the model's variables and that such uncertainty can be modelled by associating probabilities with the links between variables.

The use of BNs can capture the relation between the various nodes (variables) and consistently estimate the way in which the initial probabilities influence uncertain conclusions (such as the quality of an e-commerce system for the case at hand). In this case, BN are used for future estimation, or, forward prediction. Furthermore, BNs can be used to speculate about the states of the initial nodes, based on a given final and some intermediate variables (backward assessment).

To define the relations between the variables, we first determined the probabilities that describe the relations between a 'child' node and its 'parent' nodes. These may be collected in a Node Probability Table (NPT). For example, Figure 1 presents two 'parent' nodes (nodes B and C) and one 'child' node (node A). The probability table of node A reflects the probability P(A|B,C) for all possible combinations of A, B, C. Thus, since there are two possible states for node B (b1, b2) of Figure 1, three possible states for node C (c1, c2, c3) and three for node A (a1, a2, a3), then the NPT of node A will include 3*2*3=18 elements.

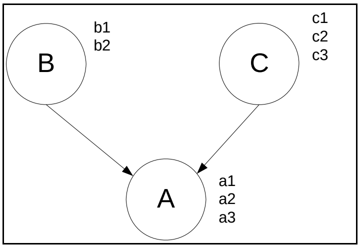


Figure 1. *A simple Bayesian network.*

4. An analysis of the quality model

Users care about quality characteristics such as functionality, reliability, usability, efficiency, flexibility, friendliness, simplicity, etc, simply because these are the characteristics which can easily be seen by the use of the product. Thus, the presented model is based on the four user-oriented quality characteristics of ISO 9126 [ISO/IEC 9126 (2001)], which are functionality, usability, reliability and efficiency. These characteristics are directly related to the quality of e-commerce systems as perceived by the end-user (but may also depend on the type of B2C e-commerce system that we study – for example, e-retailing, e-services, e-advertising, e-publishing).

Functionality [Kitchenham and Pfleeger (1996)] refers to a set of functions and specified properties that satisfy stated or implied needs. It comprises four quality sub-characteristics, which are suitability, accuracy, interoperability and security.

Usability [Fenton and Pfleeger (1997)] is defined as a set of attributes that bear on the effort needed for the use of a product or service, based on the individual assessment of such use by a stated or implied set of users. With reference to ISO 9126, usability comprises in four quality sub-characteristics: understandability, learnability, operability and attractiveness.

Efficiency [Kitchenham and Pfleeger (1996)] refers to a set of attributes that bear on the relationship between the software's performance and the amount of resources used under stated conditions. The quality sub-characteristics which efficiency comprises are time behavior, and resource behavior.

Reliability [Fenton and Pfleeger (1997)] refers to a set of attributes that bear on the capability of software to maintain its performance level under stated conditions for a stated period of time. Reliability comprises three quality sub-characteristics: maturity, fault tolerance, and recoverability.

Moving from the above high-end description towards operationally effective descriptors is work that inherently analyzes the interrelationships between various aspects of one of the core quality constituents. Such work lays the groundwork for the subsequent deployment of a BN.

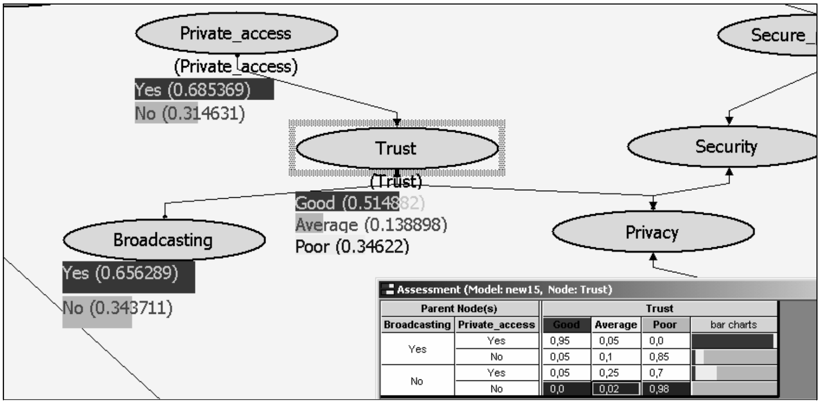
The attractiveness of an e-commerce system is based on the interface design and the alternative ways that the product or service is presented to the end-user. Fancy e-commerce web pages full of contrasting colors that seek to impress the end-user, and links that once clicked lead the end-user away from the system, are far from being good examples of usable e-commerce systems. The use of a text is the basic method of the product's presentation. The text should be short, without errors and readable, so the end-user must be able to look at the text and find all the product information he needs within a few minutes. Images and multimedia applications such us audio, video, animation and virtual reality applications lend credence to the information required. Furthermore, the existence of graphics and colors in the e-commerce systems web pages, improve the quality characteristic of attractiveness.

The learnability of the system is based on the system's template. The e-commerce systems' template should be similar with the other web pages template, so the end-user can browse in the system very easily. The quality characteristic of understandability is related to the accessibility of the e-commerce system. The system should be accessible to all users including those with mobile devices or web TV, or users with disabilities. Moreover, the understandability of the e-commerce system could be improved by services like FAQ or direct contact with the system. Finally, e-commerce systems should provide an operable searching function that offers keyword searching and other additional methods like boolean search and parametric search. The search results should be sorted along with criteria like topic, date, and compatibility and categorized according to well-organized category labels.

5. Model implementation

The overall model is composed by four sub-models. The principal node is 'Quality' and represents the overall e-commerce system quality; it is characterized by two possible states (evidence): 'good' and 'poor'. The parent nodes of 'quality' are the nodes: 'Functionality', 'Usability', 'Reliability' and 'Efficiency', namely the quality characteristics that end-users value based on ISO 9126. These quality characteristics can be also characterized by three possible states: 'good', 'average', and 'poor'. Each quality characteristic node is connected to the corresponding quality sub-characteristics, which in turn are assigned three possible states as evidence: 'good', 'average', 'poor'. Finally, each of these quality sub-characteristics is connected to intermediate nodes or to child nodes comprising the characteristics of e-commerce systems. The evidence in all intermediate nodes, that represent the characteristics of e-commerce systems, is characterized by 'good', 'average' and 'poor'. The intermediate nodes are, for example, product's presentation, searching procedure, navigation and purchase procedure. Finally, the child nodes simply answer the question posed to the user whether a specific characteristic exists in the e-commerce system or not. Aiming for a boolean verdict at the evaluator (leaf) level was a strenuously thought out decision to minimize subjectivity.

The model has been developed using the Microsoft MSBNx Authoring and Evaluation tool version 1.4.2. All the figures of the paper have been created using this tool. Each node of the model has a Node Probability Table that presents the discrete conditional probability distribution. For an example, note that in Figure 2 the intermediate node of 'Trust' is represented as a child node connected to two parent nodes, as indicated by the directed arrows. Each parent node represents the relevant e-commerce characteristics, namely: 'Private access' and 'Broadcasting'. The 'Trust' node has three possible states as evidence and the parent nodes have two states for evidence. The probability table for 'Trust' has therefore 3*2*2 = 12 elements. One of the most important characteristics affecting the successful application of the model is the definition of the Node Probability Table of each node. In the Node Probability Table of the quality sub-characteristic 'Trust', which is presented also in Figure 2, the values of the probabilities vary between 0 to 1, scaling by 0.05. The initial probabilities of the model were based on data taken from previous studies of e-commerce systems [Stefani and Xenos (2001)].



Assessment (Model: new15, Node: Trust)

Parent Node(s)		Trust			
Broadcasting	Private_access	Good	Average	Poor	bar charts
Yes	Yes	0,95	0,05	0,0	
	No	0,05	0,1	0,85	
No	Yes	0,05	0,25	0,7	
	No	0,0	0,02	0,98	

Figure 2. *Example of the model's use.*

The user can at any time insert data (evidence) for one or more nodes. This evidence can activate the conditional probabilities of other nodes and influence the corresponding estimate in a forward or backward fashion. Forward use of the model assesses the overall quality of an e-commerce system by inputting data to all base nodes and deriving an estimate about the system's quality ('good' or 'poor' with the associated probability values). Estimates are of course provided for internal model nodes as well, reflecting the quality degree of individual system components. Backward use provides assessments regarding the internal nodes, when the value of the final state of quality is defined.

In such a model, the accuracy of the predictions in terms of values measuring quality is not only related to the evidence assigned to leaves, but also highly dependent on the Node Probability Table which transforms the 'knowledge' of the model designers into probabilities. In order to refine the Node Probability Table several 'test drives' of the model were run to facilitate fine tuning.

The final test drive of the model was the assessment of 120 Business to Consumer (B2C) e-commerce systems [Stefani et al. (2004)]. These 120 systems were randomly selected from the list of numerous international e-commerce systems known at the day of this study and the evaluators assigned values to all each leaf node.

After collecting the measurement data, the next step was the analysis of the results. Firstly, the normal distribution of the data for all the quality characteristics and sub-characteristics was checked in order to ensure their validity. Consequently, the measurement data was clustered based on the systems' quality. To define these clusters, three alternative approaches could have been followed: a) assigning values that define the boundaries of the clusters a priori, i.e. before conducting any experimental measurements to the selected e-commerce systems used for the assessment, b) assigning values based on the percentages derived from the measurement results and c) assigning values based on the measurement results themselves and their possible distribution to clusters.

Although all of the above three alternatives for data analysis are acceptable, the third one was selected since it provides more representative rates. Moreover, the data analysis showed that the data were clearly distributed in different clusters. In this way we ensured that the boundaries of the different scales for each quality characteristic were more accurately defined.

6. Research directions as a case study of AI based validation

Using rules for classification purposes has been and still is one of the premier application fields of AI [Breiman et al. (1984)] [Mitchell (1997)]. The merging of numeric information into an otherwise symbolic reasoning process has also allowed us to build some grading potential into our classification tasks by allowing a verdict to be delivered in a numeric range (for example, a quality metric, as in our case) as opposed to a strictly Boolean answer.

There is no shortage of techniques to choose from if we decide to try to translate the data received from expert evaluators of commercial sites into automatic evaluation procedures. However, the approach we have started exploring has some distinct characteristics that allow more complex tasks to be carried out.

Let us first note that the ISO standard (or, any standard, for that case) defined a set of relatively high-level characteristics that an evaluator should look for in an evaluation subject. That these characteristic are relatively high level is a crucial observation. This allows them to embed within themselves several more specifying characteristics that are probably easier to define and measure. In that sense, any standard is too coarse a definition from a data collection perspective and its break-down in component characteristics is an inherent part of creating a reasonable representation for our problem task. In essence, our work first starts by selecting the features of our classification problem.

Having selected the features of our classification problem we can now embark on the data collection process. To facilitate such a process we have opted in creating features that admit a straightforward yes/no input. But we stop the data collection process at just collecting feature data; we do not assign an overall assessment to each of our measurements. The reason we do not proceed to such an assignment is that we are not yet ready to field our data collection experiment with a substantial number of experts so as to minimize the effect of individual noise into data collection. Furthermore, there is no globally accepted theory (or practice) on how the individual measurements aggregate to form an overall quality index. Note that, from a measurements perspective, the calibration process described above results in a rule-of-thumb way of ranking an e-commerce system.

It is for this reason that we have employed the BN which facilitated a preliminary theory formation on the interactions of several quality features. We acknowledge, of course, that this progress has only partially approached the goal. The open issues are quite as important.

Note that what we have actually proposed is a model of evaluation which incidentally uses a BN as its implementation platform. Is the model valid for a range of applications? To answer this question we now need to augment our experimentation with the overall assessment as well. Now, the expert evaluator will not only be invited to supply a feature-based assessment of a site; furthermore, he will be requested to supply an overall quality index.

That evaluator-supplied index is now a measure of how an individual evaluator may be “close” to our automatic evaluation policy. For sufficiently close indices, it might be reasonable to expect that the model captures the underlying practices of several individuals and can be safely deployed. However, in the initial stages of our approach we expect that we will be observing more differences than similarities.

A source of differences may be the fact that sites in different domains (for example, vertical marketplaces as opposed to horizontal ones, tangible goods as opposed to intangible ones, etc.) may warrant a different model. This would give rise to the development of domain-specific models, which cuts directly into the ontology learning research paradigm [Maedche (2002)].

A further source of differences may be the fact that some evaluators may either explicitly or implicitly treat some high-level characteristics as being composed of a different mix of components than the one we have proposed. While our proposition serves as a proof of concept, it can be readily extended to encompass multiple relationships between features as opposed to the purely hierarchical breakdown employed to-date. Identifying the weight of the individual contributions of low-level features towards the overall index is a difficult problem and, for that issue, one may have to resort first to more conventional model building approaches (for example, decision tree induction or logistic regression, even though the latter might suffer form the lack of its interpretability potential).

As a matter of fact this is the quintessential strategic aspect of machine learning in the context of evaluation: automatically building models that we will then have to analyse vis-à-vis their ad hoc human-constructed counterparts from the point of view of providing convergent reasoning paths for their results.

7. Conclusions

This paper described the development of a tool and the corresponding methodology to aid the measurement process for the quality of e-commerce systems. The measurement process is framed as a problem of probabilistic reasoning and we use a standard BN toolbox to develop the measuring tool.

Developing a measuring tool necessarily raises the question of how its applicability may be. We argue in this paper that boosting its applicability sits directly in the core of a battery of machine learning techniques. While our approach does not yet learn the graphical model from the data [Buntine (1994)], it is the particular framing of the problem that allows data collection to proceed independently of the underlying model and further allows us to independently develop new models that capture the fuzziness and subjectivity of human evaluation and compare them with our human-developed model.

Seen from that point of view, it seems that the development of a measurement methodology stands to benefit from the deployment of standard AI techniques in core points of its life-cycle. One might also speculate that this could turn out to be a valid conjecture for more software engineering fields. Interestingly, this was not so a decade ago. It is through the proliferation of new tools that the entry threshold for the exploratory application of AI techniques in several disciplines has opened up new potential for AI in the mainstream.

8. Acknowledgements

This paper shares some context-setting paragraphs with references [Stefani and Xenos (2001)] and [Stefani et al. (2004)]. However, it targets the community of AI researchers and practitioners and, hence, its focus is distinctly to the aforementioned references.

Details of the overall process of statistical analysis for the overall quality, the quality characteristics and sub-characteristics, as set out in Section 5 (along with the experimental data) can be found in the project’s web site: http://quality.eap.gr/default_en.asp.

References

Becker, S.A. and Berkemeyer, A. (2002). Rapid Application Design and Testing of Web Usability, IEEE Multimedia, vol 9, no. 4 , pp. 38-46.

Breiman, L., Friedman, J.H., Olshen, R.A. and Stone, C.J. (1984). Classification and regression trees. Belmont, CA, Wadsworth.

Buntine, W. (1994). Operations for learning with graphical models, Journal of Artificial Intelligence Research, vol. 2, pp. 159-225.

Burrows, J. (1999). Information Technology standards in a changing world: the role of the users, Computer standards & Interfaces, vol. 20, pp. 323 – 331.

Chang, W.K., and Hon, S.K. (2002). Assessing the Quality of Web-Based Applications via Navigational Structures, IEEE Multimedia, vol. 9, no. 3, pp. 22-30.

Fenton, N. and Pfleeger, S. (1997). Software Metrics A Rigorous & Practical Approach, Thomson Computer Press.

ISO/IEC 9126 (2001). Software Product Evaluation – Quality Characteristics and Guidelines for the User, International Organization for Standardization, Geneva.

Ivory, M.Y. and Hearst, M.A. (2003). Improving Web Site Design, IEEE Internet Computing, vol. 6, no. 2, pp. 56-63.

Kitchenham, B. and Pfleeger, S. (1996). Software Quality: The Elusive Target, IEEE Software, vol 13, no. 1, pp. 12-21.

Lauritzen, S. and Spiegelhalter, J. (1990). Local computations with probabilities on graphical structures and their applications to expert systems, Readings in Uncertain Reasoning, San Mateo, California, pp. 415 – 448.

Maedche, A. (2002). Ontology Learning for the Semantic Web. Springer.

Mich, L., Franch, M., and Gaio, L. (2003). Evaluating and Designing Web Site Quality, IEEE Multimedia, vol. 10, no. 1, pp. 34-55.

Mitchell, T. (1997). Machine Learning. McGraw Hill.

Pearl, J. (1986). Fusion, propagation, and structuring in belief networks, Artificial Intelligence, vol. 29, no. 3, pp. 241-288.

Stefani, A., Stavrinoudis, D., and Xenos, M. (2004). Experimental Based Tool Calibration used for Assessing the Quality of E-Commerce Systems, Proceedings of the 1st IEEE International Conference on E-Business and Telecommunication Networks (ICETE-2004), Portugal, Vol. 1, pp. 26- 32.

Stefani, A. and Xenos, M. (2001). A model for assessing the quality of e-commerce systems, Proceedings of the Panhellenic Conference with International Participation in Human Computer Interaction (PC-HCI 2001), pp. 105- 109.

Zwass, V. (1996). Electronic commerce: structures and issues, International Journal of Electronic Commerce, vol. 1, pp. 3– 23.