From Web Pages to Web Communities

Miloš Kudělka[1], Václav Snášel[1], Zdeněk Horák[1], and Aboul Ella Hassanien[2]

[1] VSB Technical University Ostrava, Czech Republic
milos.kudelka@inflex.cz, {vaclav.snasel, zdenek.horak.st4}@vsb.cz
[2] Faculty of Computer and Information, Information Technology Department, Cairo University, Egypt
abo@cba.edu.kw

Abstract. In this paper we are looking for a relationship between the intent of Web pages, their architecture and the communities who take part in their usage and creation. From our point of view, the Web page is entity carrying information about these communities and this paper describes techniques, which can be used to extract mentioned information as well as tools usable in analysis of these information. Information about communities could be used in several ways thanks to our approach. Finally we present an experiment which illustrates the benefits of our approach.

Keywords: Web community, Web site, Web pattern, genre

1 Introduction

Metaphor: A Web page is like a family house. Each of its parts has its sense, determined by a purpose which it serves. Every part can be named so that everybody imagines approximately the same thing under that name (living room, bathroom, lobby, bedroom, kitchen, balcony). In order that the inhabitants may orientate well in the house, certain rules are kept. From the point of view of these rules, all houses are similar. That is why it is usually not a problem e.g. for first time visitors to orientate in the house. We can describe the house quite precisely thanks to names. If we add information about a more detailed location such as sizes, colors, equipment and further details to the description, then the future visitor can get an almost perfect notion of what he will see in the house when he comes in for the first time. We can also approach the description of a building other than a family house (school, supermarket, office etc.). Also in this case the same applies for visitors and it is usually not a problem to orientate (of course it does not always have to be the case, as well as bad Web pages there are also bad buildings).

In the case of buildings, we can naturally define three groups of people, which are somehow involved in the course of events. The first group are the people defining the intent and the purpose (those who pay and later expect some profit), the second one are those who construct the building (and are getting paid for it)

K. Richta, J. Pokorný, V. Snášel (Eds.): Dateso 2009, pp. 13–22, ISBN 978-80-01-04323-3.

and the third group are "users" of the building. These groups fade into another and change as society and technology evolve.

As we describe in the subsequent text, the presented metaphor can - up to certain point - serve as an inspiration to seize the Web pages content and also the whole Web environment.

This text is organized as follows. In the second section we describe the Web page from the view of groups of people sharing the Web page existence. The third section describes tools and techniques required for our experiment. In particular our own Pattrio method, which is designed to detect Design patterns within Web pages, and FCA used for clustering. In the fourth section we describe experiment dealing with Web site description. The last section contains paper recapitulation and focuses on possible directions of further research.

2 From Web pages to Web communities

Every single Web page (or group of Web pages) can be perceived from three different points of view. When considering the individual points of view we were inspired by specialists on Web design ([29]) and on the communication of humans with computers ([6]). These points of view represent the views of three different groups of communities who take part in the formation of the Web page (fig. 1).

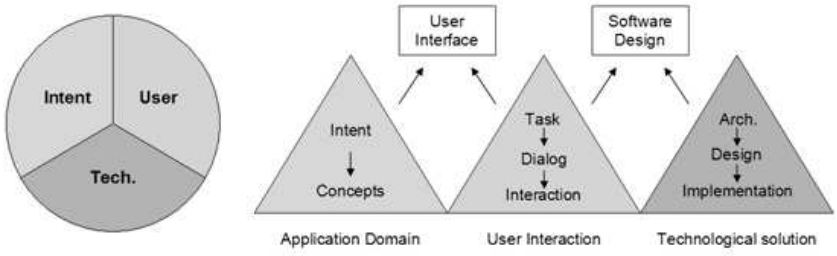


Fig. 1. Views of three different groups

(1) The first group are those whose intention is that the user finds what he expects on the Web page. The intention which the Web page is supposed to fulfill is consequently represented by this group. For the sake of clarity, we can say that this group is often represented by Web site owners. (2) The second group are developers responsible for the creation of the Web page. They are therefore consequently responsible for fulfilling the goals of the two remaining groups. (3) The third group are users who work with the Web page. This group consequently represents how the Web page should appear outwardly to the user. It is important that this performance satisfies a particular need of the user.

As an example we can mention blogs. The first community are the companies, which offer an environment and technological background for blog authors and to

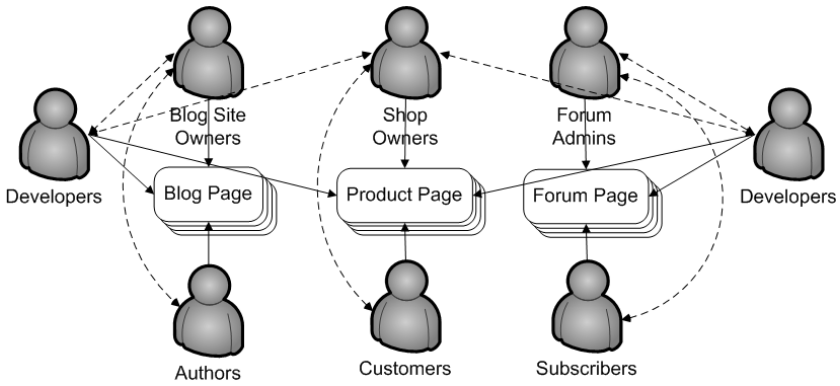


Fig. 2. Social network around Web pages

some extent they also define the formal aspects of blogs. The second community are the developers who implement the task given by the previous group. The visible attribute of this group is that they – to a certain degree – share their techniques and policies. The third group consists of blog authors (in the sense of content creation). They influence the previous two groups retroactively. The second example can be the product pages - the intention of the e-shop is to sell items (concretely to have Web pages where you can find and buy the products), the intention of the developers is to satisfy the e-shop owners as well as the Web page visitors. The intention of the visitors is to buy products, so they expect clearly stated and well-defined functionality. From this point of view, the web pages are elements around which the social networks are formed (fig. 2). For further details and references, please see [1] and [15] (which considers also the aspect of network evolution).

Under the term *Web community* we usually think of a group of related Web pages, sharing some common interests (see [28], [20], [21]). As a Web community we may also consider Web site or groups of Web sites, on which people with common interests interact. It is apparent, that all three aforementioned groups participate in the Web page life cycle. The evolution of a page is directly or indirectly controlled by these groups. As a consequence, we can understand the Web page as a projection of interaction among these three groups. The analysis of the page content may uncover significant information, which can be used to assign the Web page to a Web community.

3 Tools and techniques

Our aim is to automatically discover such information about Web pages, that comes out of intentions of particular groups. Using these information we can find the relations between the communities and describe them (on the technical level). The key element for Web page description is the name of the object,

which represents the intention of the page. It can be “Home page”, “Blog” or “Product Page”. In the detailed description we can distinguish, for example, between “Discussion”, “Article” or “Technical Features”. We can also use more general description, such as “Something to Read” or “Menu” (see [14]).

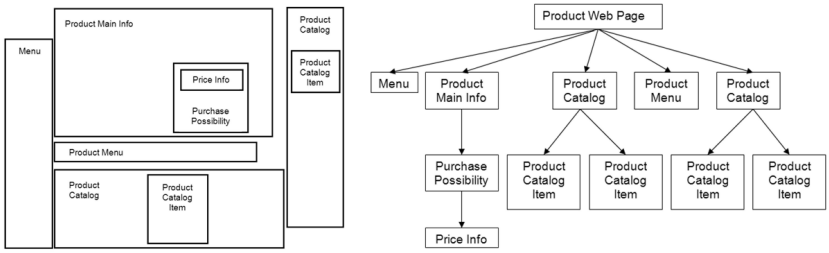


Fig. 3. Product page scheme (a), (b)

The first group of intentions represents so-called Genre (see [5]). The second group is very close to Web Design Patterns [30]. Figure 3 contains schematically depicted product Web page with hierarchy of solved tasks (each task represents one particular intention). The ability to discover aforementioned elements (Genres and Web design patterns) is required to obtain the Web page description (and consequently also the intentions represented by mentioned communities).

Genre is a taxonomy that incorporates the style, form and content of a document which is orthogonal to topic, with fuzzy classification to multiple genres [4]. In the same paper are described existing classifications. Regarding these classifications there are many approaches on genre identification methods. The goal of paper [11] is to analyze home page genres (personal home page, corporate home page or organization home page). In paper [7] authors have proposed a flexible approach for Web page genre categorization. Flexibility means that the approach assigns a document to all predefined genres with different weights. In [9] paper, there is described a set of experiments to examine the effect of various attributes of Web genre on the automatic identification of the genre of Web pages. Four different genres are used in the data set (FAQ, News, E-Shopping and Personal Home Pages).

3.1 Pattrio method

Design patterns describe proven experience of repeated problem solving in the area of software solution design. While the design patterns have been proven in real projects, their usage increases the solution quality and reduces the time of their implementation. Good examples are also the so called Web design patterns, which are patterns for design related to the Web. Even in this area, the patterns are getting quite common (they are collected and published in the form of printed or Internet catalogues, e.g. see [29], [30]).

We have designed our own Pattrio method used for the detection of Web design pattern instance in web pages. In Pattrio method we work with 24 patterns (mostly e-commerce and social domain). Pattrio method is based on analysis of technical (architectural) and semantical attributes of solutions of the same tasks in the environment of Web, for details see [13], [14].

Detection algorithm In the context of our approach, there are elements with semantic contents (words or simple phrases and data types) and elements with importance for the structure of the web page where the Web pattern instance can be found (technical elements). The rules are the way that individual elements take part in the Web pattern display. While defining these rules, we have been inspired by the Gestalt principles (see [27]). We are using four rules based on these principles. The first one (proximity) defines the acceptable measurable distances of individual elements from each other. The second one (closure) defines the way of creating of independent closed segments containing the elements. One or more segments then create the Web pattern instance on the web page. The third one (similarity) defines that the Web pattern includes more related similar segments. The forth one (continuity) defines that the Web pattern contains more various segments that together create the Web pattern instance. The relations among Web patterns can be on various levels similar as classes in OOP (especially simple association and aggregation).

The basic algorithm for detection of Web patterns then implements the preprocessing of the code of the HTML page (only selected elements are preserved e.g. block elements as table, div, lines, etc.), segmentation and evaluation of rules and associations. The result for the page is the score of Web patterns that are present on the page. The score then says what is the relevance of expecting the Web pattern instance on the page for the user.

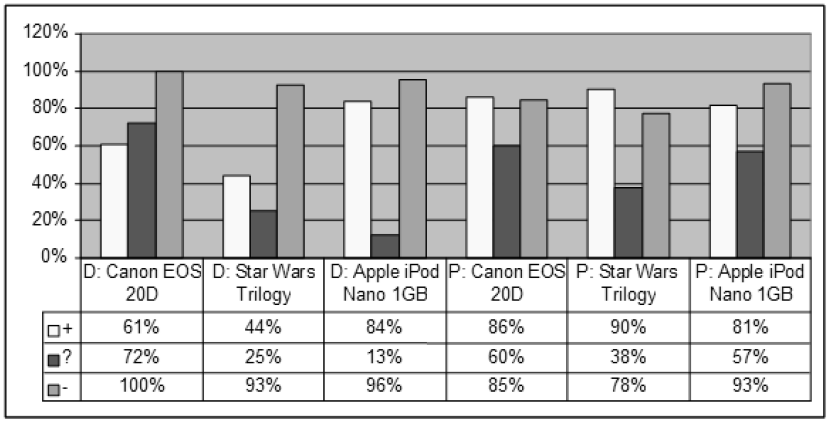
The accuracy of our method is about 80% (see [12]). Figure 4 shows the accuracy of Pattrio method for tree selected products (Apple iPod Nano 1GB, Canon EOS 20D, Star Wars Trilogy film) and for the *Discussion* pattern and the *Purchase possibility* pattern. We used only the first 100 pages for each product. We manually and using Pattrio method evaluated the pages using a three-degree scale:

+ Page contains required pattern.
? Unable to evaluate results.
- Page do not contain required pattern.

Then we compared these evaluations. For example the first value 61% expresses the method accuracy for the pages with Canon EOS 20D product where there was a discussion.

3.2 Formal Concept Analysis

As one of the suitable tools for analyzing this kind of data we consider Formal concept analysis. When preprocessing Web pages we often cannot clearly state



	D: Canon EOS 20D	D: Star Wars Trilogy	D: Apple iPod Nano 1GB	P: Canon EOS 20D	P: Star Wars Trilogy	P: Apple iPod Nano 1GB
□+	61%	44%	84%	86%	90%	81%
■?	72%	25%	13%	60%	38%	57%
■-	100%	93%	96%	85%	78%	93%

Fig. 4. Accuracy of Pattrio method for detection of *Discussion* and *Purchase Possibility* patterns - percentage of agreement between human and Pattrio method evaluation on sets of Web pages returned for different search queries

the presence of an object in the page content. We are able to describe the amount of its presence at some scale and this information can be captured using fuzzy methods and analyzed using a fuzzy extension of Formal Concept Analysis ([3]). But since we are dealing with a large volume of data ([8]) and a very imprecise environment, we should consider several practical issues, which have to be solved prior the first applications. Methods of matrix decomposition have succeeded in reducing the dimensions of input data (see [26] for application connected with Formal concept analysis and [18], [17] for overview).

Formal concept analysis (shortly FCA, introduced by **Rudolf Wille** in 1980) is well known method for object-attribute data analysis. The input data for FCA we call **formal context** C, which can be described as $C = (G, M, I)$ – a triplet consisting of a set of objects G and set of attributes M, with I as relation of G and M. The elements of G are defined as objects and the elements of M as attributes of the context.

For a set $A \subseteq G$ of objects we define $A^{'}$ as the set of attributes common to the objects in A. Correspondingly, for a set $B \subseteq M$ of attributes we define $B^{'}$ as the set of objects which have all attributes in B. A **formal concept** of the context (G, M, I) is a pair (A, B) with $A \subseteq G$, $B \subseteq M$, $A^{'} = B$ and $B^{'} = A$. $\mathcal{B}(G, M, I)$ denotes the set of all concepts of context (G, M, I) and forms a complete lattice (so called **Gallois lattice**). For more details, see [10].

4 Experiment

For the need of our experiment we have implemented a Web application with user interface connected to the API of different search engines (google.com, msn.com, yahoo.com and the Czech search engine jyxo.cz above all). Users from a group

of students and teachers of high schools and our university were using this application for more than one year to search for ordinary information. We have not influenced the process of searching in any way. The purpose of this part of experiment was to view the World Wide Web using the perspective of normal users (as the search engines play key role in World Wide Web navigation). In the end we have obtained dataset with more than 115,000 Web pages. After clean up, 77,850 unique Czech pages remained. For every single Web page we have performed the detection of sixteen objects. The page did not have to contain any object, as well as it may have contained 16 objects (Price information, Purchase possibility, Special offer, Hire sale, Second hand, Discussion and comments, Review and opinion, Technical features, News, Enquire, Login, Something to read, Link group, Price per item, Date per item, Unit per item). We have used such preprocessed dataset for following experiment.

In the experiment we have tried to visualize the structure and relations of Web sites (and as a result also Web communities) referring to one specific topic. As an input we have used the list of domains created in the previous experiment. Only Web sites with more than 20 pages in the dataset have been taken into consideration. Each domain is accompanied by detected objects. This list is transformed into a binary matrix and considered as a formal context. Using methods of FCA we have computed a concept lattice which can be seen on figure 5. The resulting matrix has 516 rows (objects) and 16 columns (attributes), computed concept lattice contains 378 concepts.

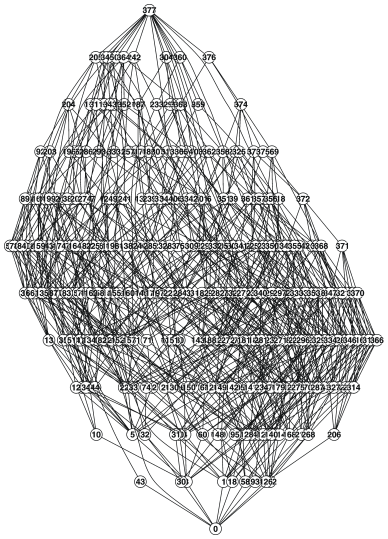


Fig. 5. Lattice calculated from whole dataset

From the computed lattice we have selected a sub-lattice containing 18 Web sites dealing with cell phones. Only 5 attributes have been selected and the visualization was created in a slightly different manner (see figure 6 and attached legend). Each node of the graph corresponds to one formal concept. To increase the visualization value, the attributes are represented by icons and the set of objects (Web sites) is depicted using small filled/empty squares in the lower part. It can be easily seen that using created visualization we can think of dividing the whole set of Web sites into two groups - the first one contains sites where users are enabled to buy cell phones and the second one where the users are allowed to have a discussion. The illustrated division is in the soft sense only — one may think of concept nr. 8 as being part of the shopping group also. Web sites presented in higher levels of lattice are considered in more specific context. Deeper insight gives you more detailed information about Web site structures and relations.

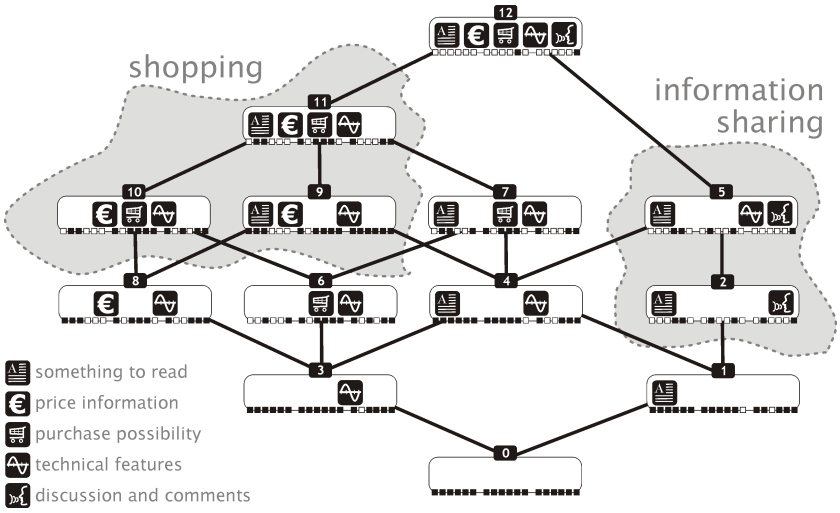


Fig. 6. Part of lattice

The concept lattice forms a graph, which can be interpreted as an expression of relation between different Web sites. As a consequence, it describes the relation between different Web communities, because behind the shopping-related domains we can see the group of users interested in buying cell phones and behind the information–sharing pages we see the community of users interested only in the technical aspects, features of cell phones and their discussing.

5 Conclusions and future work

In this paper we have described three kinds of social groups which take part in Web page creation and usage. We distinguish these groups using their relation to the Web page - whether they define the intent of the page, whether they create the page or whether they use the page. By using this analysis we can follow the evolution of the communities and observe the expectancies, rules and behavior they share. Obtained information can be surely used to improve the searching process. From this point of view, Web 2.0 is only a result of the existence and interaction of these social groups.

Our experiment shows that if we focus ourselves on Web sites and the Web page content they provide, we can ask interesting questions. These questions may bear upon the Web sites' similarity and the similarity of social groups involved with these pages. For us this shows the direction of further research in which we will investigate answers to these questions in more detail.

References

1. L. Adamic, E. Adar: How to search a social network, Journal Social Networks, vol. 27, pp. 187–203 (2005)
2. Ch. Alexander: A Pattern Language: Towns, Buildings, Construction, Oxford University Press, New York (1977)
3. R. Belohlavek, V. Vychodil: What is a fuzzy concept lattice, Proceedings of the CLA, 3rd Int. Conference on Concept Lattices and Their Applications, pp. 34–45 (2005)
4. E. S. Boese: Stereotyping the web: Genre classification of Web documents, Colorado State University (2005)
5. E. S. Boese, A. E. Howe: Effects of web document evolution on genre classification, Proceedings of the 14th ACM international conference on Information and knowledge management, pp. 632–639 (2005)
6. J. O. Borchers: Interaction Design Patterns: Twelve Theses, Workshop, The Hague, vol. 2 (2000)
7. J. Chaker, O. Habib: Genre Categorization of Web Pages, Proceedings of the Seventh IEEE International Conference on Data Mining Workshops, pp. 455–464 (2007)
8. R. J. Cole, P. W. Eklund: Scalability in Formal Concept Analysis, Computational Intelligence, vol. 15, pp. 11–27 (1999)
9. L.Dong, C. Watters, J. Duffy, M. Shepherd: An Examination of Genre Attributes for Web Page Classification, Proceedings of the Proceedings of the 41st Annual Hawaii International Conference on System Sciences, pp. 133–143 (2008)
10. B. Ganter, R. Wille: Formal Concept Analysis: Mathematical Foundations, Springer-Verlag, New York (1997)
11. A. Kennedy, M. Shepherd: Automatic Identification of Home Pages on the Web, Proceedings of the 38th Hawaii International Conference on System Sciences (2005)
12. J. Kocibova, K. Klos, O. Lehecka, M. Kudelka, V. Snasel: Web Page Analysis: Experiments Based on Discussion and Purchase Web Patterns, Web Intelligence and Intelligent Agent Technology Workshops, pp. 221–225 (2007)

13. M. Kudelka, V. Snasel, O. Lehecka, E. El-Qawasmeh: Semantic Analysis of Web Pages Using Web Patterns, Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence, pp. 329–333 (2006)
14. M. Kudelka, V. Snasel, O. Lehecka, E. El-Qawasmeh, J. Pokorny: Web Pages Reordering and Clustering Based on Web Patterns, SOFSEM 2008, pp. 731–742 (2008)
15. R. Kumar, J. Novak, A. Tomkins: Structure and Evolution of Online Social Networks, Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 611–617 (2006)
16. D. Lee, O. R. Jeong, S. Lee: Opinion mining of customer feedback data on the web, Proceedings of the 2nd international conference on Ubiquitous information management and communication, pp. 230–235 (2008)
17. D. Lee, H. Seung.: Learning the parts of objects by non-negative matrix factorization, Nature, vol. 401, pp. 788–791 (1999)
18. T. Letsche, M. W. Berry, S. T. Dumais.: Computational methods for intelligent information access, Proceedings of the 1995 ACM/IEEE Supercomputing Conference (1995)
19. H. Y. Limanto, N. N. Giang, V. T. Trung, J. Zhang, Q. He, N. Q. Huy: An information extraction engine for web discussion forums, International World Wide Web Conference, pp. 978–979 (2005)
20. T. Murata: Discovery of User Communities from Web Audience Measurement Data, Web Intelligence 2004, pp. 673–676 (2004)
21. T. Murata, K. Takeichi: Discovering and Visualizing Network Communities, Web Intelligence/IAT Workshops 2007, pp. 217–220 (2007)
22. Z. Nie, J. R. Wen, W. Y. Ma: Object-level Vertical Search, Third Biennial Conference on Innovative Data Systems Research, pp. 235–246 (2007)
23. Z. Pawlak: Rough Sets: Theoretical Aspects of Reasoning about Data, Kluwer Academic Publishing (1991)
24. M. A. Rosso: User-based identification of Web genres. JASIST (JASIS) 59(7), pp. 1053–1072 (2008)
25. S. Schmidt, H. Stoyan: Web-based Extraction of Technical Features of Products, Beiträge der 35. Jahrestagung der Gesellschaft für Informatik, pp. 256–261 (2005)
26. V. Snasel, M. Polovincak, H. M. Dahwa, Z. Horak: On concept lattices and implication bases from reduced contexts, Supplementary Proceedings of the 16th International Conference on Conceptual Structures, ICCS 2008, pp. 83–90 (2008)
27. J. Tidwell: Designing Interfaces: Patterns for Effective Interaction Design, O'Reilly, pp. 0–596 (2005)
28. M. Toyoda, M. Kitsuregawa: Creating a Web community chart for navigating related communities, Hypertext 2001, pp. 103–112 (2001)
29. D. K. Van Duyne, J. A. Landay, J. I. Hong: The Design of Sites: Patterns, Principles, and Processes for Crafting a Customer-Centered Web Experience, Addison-Wesley Professional (2003)
30. M. Van Welie: Pattern Library for Interaction Design. www.welie.com, (last access 2008-08-07)
31. S. Zheng, R. Song, J. R. Wen: Template-independent news extraction based on visual consistency, In Proceedings of the Twenty-Second AAAI Conference on Artificial Intelligence, pp. 1507–1513 (2007)