Abstract—This paper proposes a novel Web site description that can be used for simple visualization and reasoning about Web sites similarity from the viewpoint of users, developers and owners. For this description, we use objects closely related to Genres and Web design patterns. We detect these objects on single page using a novel approach called Pattrio method. Web site description is obtained from the frequency of these objects on individual pages. Some experiment results illustrating the usability of the approach and possible direction for further research is also included.

Keywords—genre; design pattern; web community; web site;

I. INTRODUCTION

Web is a product of permanent interaction between different social groups. The cornerstone of the Web is the Web page, which is, as a consequence, projection of this interaction. The data containing behaviour of users are one of the results of this interaction in the Web environment. Such data can be used (but also misused) for various purposes. The proposed framework deals with only Web page contents. Information about cooperation of mentioned social groups can be extracted from the Web page content. Using the Web page content analysis of concrete Web sites, we illustrate what they have in common, focused groups and various interesting associated tasks.

Metaphor: A Web page is like a family house. Each part 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 get oriented (of course it does not always have to be the case, since there are also bad buildings and poorly designed Web pages).

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) 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.

Remaining text is organized as follows. In the following 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 the experiments. In particular the Pattrio method, which is designed to detect Design patterns within Web pages, and FCA used for clustering. In the Fourth Section, we describe experiments dealing with Web site description followed by conclusions and possible directions of further research.

II. FROM WEB PAGES TO WEB COMMUNITIES

Every single Web page (or group of Web pages) can be perceived from three different point of view. When considering the individual point of view, we were inspired by specialists on Web design ([16]) and on the communication of humans with computers ([4]). These aspects represent the views of three different groups of communities who take part in the formation of the Web page (fig. 1).

(1) The first group are those whose intention is that the user finds what he expects on the Web page. The intention for which the Web page is supposed to fulfill is consequently represented by this group. For the sake of clarity, we may call 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 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). Readers may consult [1] and [11], which also considers the aspect of network evolution.

Figure 1. Views of three different groups

Figure 2. Social network around Web pages

III. COMPUTATIONAL 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, one can find the relations between the communities and describe them technically. 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 a more general description, such as “Something to Read” or “Menu” (see [10]).

A. Genres and Web Design Patterns

The first group of intentions represents so-called Genre ([3]) and the second group is very close to Web design patterns [17]. Figure 3 contains schematically depicted product Web page with an 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 [2]. Regarding these classifications, there are many approaches for genre identification methods. The goal of the paper [7] is to analyze home page genres (personal home page, corporate home page or organization home page). Authors in [5] 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. Authors in [6], described a set of experiments to examine the effect of various attributes of 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).

In the following text, we use the term Named object for Genres and Web design patterns.

B. 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 [16], [17]).

We developed the Pattrio method for the detection of Web design pattern instances in web pages. In the Pattrio method, we deal with 24 patterns (mostly e-commerce and social domain). Pattrio method is based on the analysis of technical (architectural) and semantical attributes of solutions of the same tasks in the environment of Web. For technical details, reader may consult [9] and [10].

1) Detection Algorithm: In the context of the proposed 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 Figures 4 and [14]). We formulated four rules based on these principles. The first rule (proximity) defines the acceptable and measurable distances of individual elements from each other. The second rule (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 rule (similarity) defines that the Web pattern includes more related similar segments. The forth rule (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 pre-processing of the code of the HTML page (only selected elements are preserved e.g. block of elements as tables, 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 entire process including Web design pattern detection is displayed in Figure 5.

The accuracy of the proposed method is about 80% ([8]). Figure 6 illustrates the accuracy of Pattrio method for tree selected products (Apple iPod Nano 1GB, Canon EOS 20D, Star Wars Trilogy film). Discussion pattern and the Purchase possibility pattern. We used only the first 100 pages for each product. For comparison purposes, we used Pattrio method and manual evaluation of the Web pages using a three-degree scale given below:

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

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

IV. EXPERIMENTS

In this paper, we attempt to find such a description of a Web site that comes from the analysis of the architecture of pages belonging to this Web site. We treat the
Named objects as the basis of this architecture, because they encapsulate those parts of the page content that have partial intent. Named objects are closely related to so-called Genres and we used the Pattrio method for their identification. Our Web site description indicates the intent and purpose of the Web site. From this point of view, the description provides interesting information about Web sites and as a consequence, we consider it as a description of the Web community.

As a Web site, we consider a collection of Web pages placed together on one or more servers available via Internet. Web pages from one Web site share the same URL prefix and link to themselves. URL addresses of individual pages are organized into a hierarchy, which allow users to orient themselves in a Web site. The root of this hierarchy is usually special, a Web page known as the Home page. From a technical point of view, we understand a Web site as an Internet domain. This view can be inaccurate on a certain level, especially for Internet domains providing Web hosting. However, such domains are also usually specialized in some way, e.g. blogs, corporate and personal pages or small e-shops.

Various Web sites exist for various reasons. As typical examples, we can consider e-shops, news servers, social-related Web sites, corporate and academic Web sites, personal Web sites, etc. Web sites have different content and size. For example personal Web site can contain a small collection of Web pages, but a social-related Web site can have more than a million of Web pages.

From an external view, one Web site can appear differently to different users. Also the reasons to visit the Web site may vary. Let’s take an e-shop as an example. It is sure that the aim of the e-shop owner is to sell the goods. The aims of visitors may vary. A user can visit the e-shop to explore the kinds and prices of goods and read the terms of sale. The main target is the price information and maybe the price comparison. Another user wants to directly buy the goods. In that case, the visitor will be interested in pages with a purchasing possibility. A third user may be interested in product parameters and the opinions of other users. Therefore he will prefer pages containing technical features of products, discussion, FAQ, customer reviews and ratings. Web developers may have also another goal. Successful solutions and typically used compositions appear on Web pages in different Web sites. Therefore some kind of unification could be seen in the development of Web pages with usual intent. This unification is based on principles which comes out of simple consideration: “Let’s do the things like others do successfully.” Developers can follow the progress of their competitors. They can incorporate the new and successful techniques they have seen at their rivals’.

It is hard to imagine that in the era of Internet search engines, users would always search Web pages by direct visit and navigation from a Home page. One expects them to use one of the search engines to find the required page. In that case, they will probably avoid the Home page and will be navigating only through a limited part of the Web site, offered by the search engine in the first step. On the other hand, they can miss some pages completely.

For experiment purposes, 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 VSB Technical University Ostrava were using this application for more than one year to search for ordinary information. We have not influenced the search process in any way. In the end, we 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 Named objects. The page did not have to contain any Named object, as well as it may have contained 16 Named 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 the preprocessed dataset for experiments reported in this paper.

For visualization of extracted information, we have adopted a common illustration method, which works as follows. In the center of figure 7 the circle representing Web site is depicted. The size of the circle is determined by the relative size (number of pages) in the dataset. The circles around the Web site correspond to individual Named objects. Size of circles is again determined by their relative presence in the dataset.

Web site is connected to Named objects using straight line. The strength of this connection (represented by the line width) corresponds to the detected degree of Named object presence in Web site.

Figures 7, 8, 9 contain the described visualization of some Web sites from Czech Republic, which are typical Web sites.
Figure 7. Typical Web site aimed at selling products

Figure 8. Typical Web site aimed at information sharing

Figure 9. Typical university Web site
aimed at selling products, sharing information and education.

Figure 10 depicts two Web sites with similar intent - selling products. The second one differs in allowing users to share information in addition to the purchase possibility.

Presented views on Web sites allows comprehensive insight into the Web site essence. As a consequence, it also allows us to measure the similarity of Web sites. Similar Web sites can be understood as members of the same community.

V. CONCLUSIONS AND FUTURE WORK

In this paper, we have described three kinds of social groups which take part in the Web page creation and usage. We distinguish these groups by 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 are able to follow the evolution of the Web communities and observe the expectancies, rules and behavior they share. Obtained information can be surely used to improve the search process. From this point of view, Web 2.0 is only a result of the existence and interaction of these social groups.

Our experiments illustrate that Web sites and the Web page content they provide, provide interesting research questions. These questions may bear upon the Web sites’ similarity and the similarity of social groups involved with these pages. All these certainly provides the direction of further research where we will investigate answers to some of these difficult questions.

REFERENCES


