Chapter 13
Web Communities Defined by Web Page Content

Miloš Kudělka, Václav Snášel, Zdeněk Horák, Aboul Ella Hassanien, and Ajith Abraham

Abstract In this chapter, 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. For us, the Web page is entity carrying information around these communities. Our chapter describes techniques which can be used to extract the mentioned information as well as tools usable in the analysis of this information. Information about communities could be used in several ways thanks to our approach. Finally, we present experiments which prove the feasibility of our approach. These experiments also show a possible way as to how to measure the similarity of Web pages and Web sites using microgenres. We define the microgenre as a building block of Web pages which is based on the social interaction.

13.1 Introduction

The current Web – due its dimensions, number of Web pages and Web sites – starts to exceed the scope of human perception. The problem with orientation in the Web space is a consequence of this. As evidenced, we can consider the problems of contemporary search engines with giving the relevant answers to user queries. Therefore, new tools are still arising, and their authors come with new approaches supporting interaction between users and computers in the Internet environment. On one hand, in our chapter, we traverse through the field of Web content mining.

M. Kudělka (✉), V. Snášel, and Z. Horák
VSB Technical University Ostrava, Czech Republic
e-mail: milos.kudelka@inflex.cz; [vaclav.snasel;zdenek.horak.st4]@vsb.cz

A.E. Hassanien
Faculty of Computer and Information, Information Technology Department,
Cairo University, Egypt
e-mail: abo@cba.edu.kw

A. Abraham
Machine Intelligence Research Labs (MIR Labs), Scientific Network for Innovation and Research Excellence, USA
e-mail: ajith.abraham@ieee.org

A. Abraham et al. (eds.), *Computational Social Network Analysis*, Computer Communications and Networks, DOI 10.1007/978-1-84882-229-0_13,

On the other hand, we try to consider the content mined from Web pages as a result of interaction between different social groups of people in the Web environment.

Metaphor A Web page is like a family house. Each of its parts has its purpose, determined by a function which it serves. Every part can be named so that all users envision approximately the same thing under that name (living room, bathroom, lobby, bedroom, kitchen, and balcony). In order for the inhabitants to 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 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, furnishings 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 or she comes in for the first time. We can also take an approach similar to 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 there are bad Web pages, there are also bad buildings).

In the case of buildings, we can naturally define three groups of people, who are somehow involved in the course of events. The first group is the people defining the intent and the purpose (those who pay and later expect some profit), the second group is those who construct the building (and are getting paid for it), and the third group is "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. We also define the notion of the microgenre as a buliding block of Web page. The name of microgenre on Web page can serve as a linking element between groups of people with different intentions. In the third section, we provide an overview of related approaches and methods mostly from the area of Web genres identification and detection, Web design patterns and information extraction methods. The fourth section describes tools and techniques required for our experiments. In particular, our own Pattrio method, which is designed to detect microgenres within Web pages, and FCA are used for clustering. In the fifth section, we describe two experiments dealing with Web site description. The last section contains chapter recapitulation and focuses on possible directions of further research.

13.2 From Web Pages to Web Communities

The situation is similar with Web pages and 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

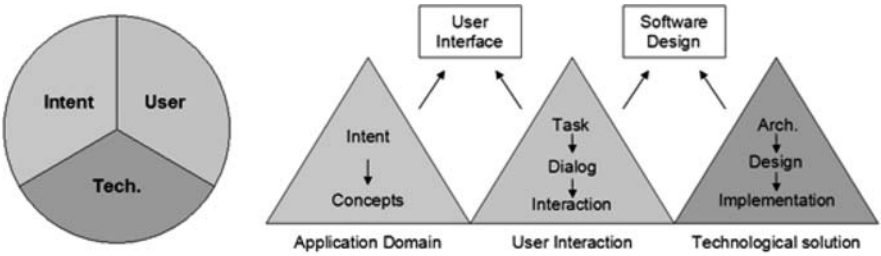


Fig. 13.1 Views of three different groups

design [33] and the communication of humans with computers [5]. These points of view represent the views of three different groups of communities who take part in the formation of the Web page (Fig. 13.1).

(1) The first group consists of 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. (2) The second group consists of developers responsible for the creation of the Web page. They are therefore consequently responsible for fulfilling the goals of the two other groups. (3) The third group consists of 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 consists of the companies that offer the environment and the technological background for blog authors and to some extent also define the formal aspects of blogs. The second community consists of 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), and 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, and 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. 13.2). For further details and references, please see [1, 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 [22, 23, 32]). As a Web community we may also consider a 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 the 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.

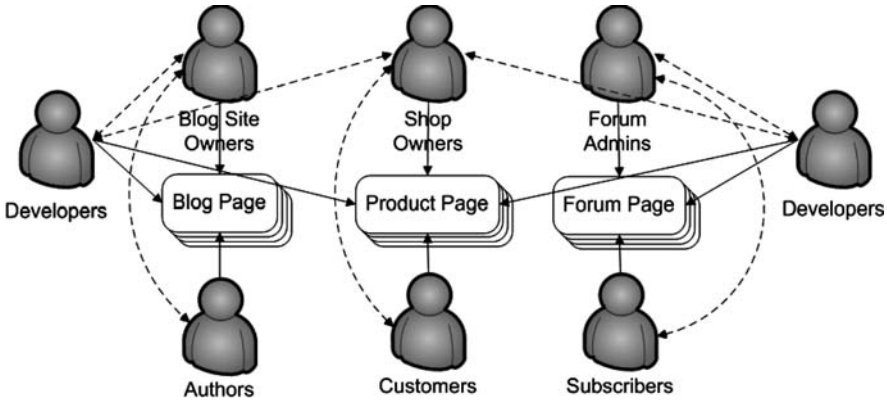


Fig. 13.2 Social network around Web pages

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 or part 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]).

13.2.1 Microgenres

According to Wikipedia.org, the genre is the division of concrete forms of art using the criteria relevant to the given form (e.g. film genre, music genre and literature genre). In all sectors of the arts, the genres are vague categories without fixed boundaries and are especially formed by the sets of conventions.

Many artworks are cross-genre and employ and combine these conventions. Probably the most deeply theoretically studied genres are the literary ones. It allows us to systematize the world of literature and consider it as a subject of scientific examination. We can find the term microgenre in this field. For example, in [21] the microgenre is seen as part of a combined text. This term has been introduced to identify the contribution of inserted genres to the overall organization of text. The motivation to use the term "microgenre" is because it is used as a building block of the analytic descriptive system. On the other hand, Web design patterns are used more technically and provide means for good solution of Web pages.

From our point of view the Web page is structured similarly to the literature text using parts which are relatively independent and have their own purpose (see Fig. 13.3). On the other side, the architecture of the Web page (individual parts)

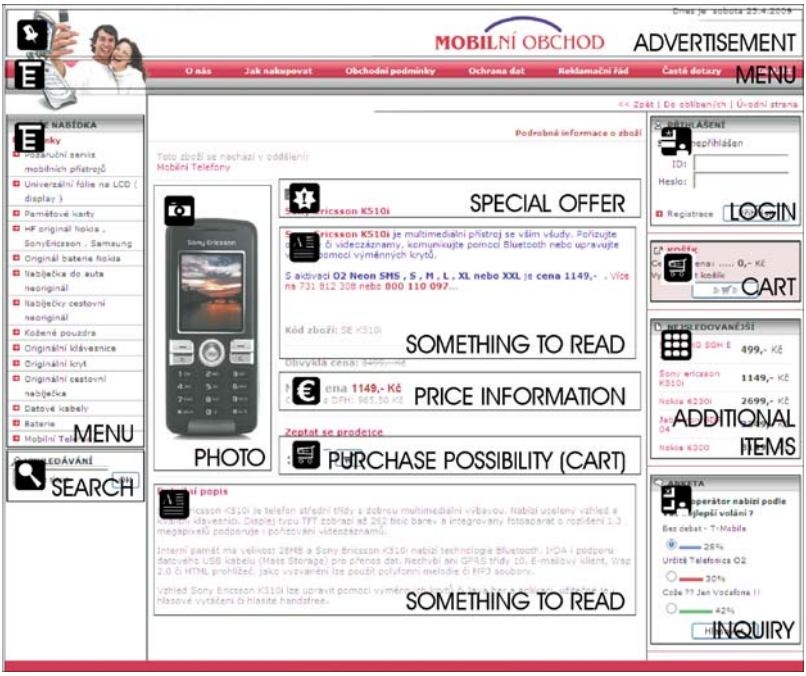


Fig. 13.3 Structure of Web page

is based on Web design patterns. For these parts, we have chosen the term "microgenre". Contemporary Web pages are often very complex, nevertheless they can usually be described using several microgenres. This kind of description can be more flexible than the genre description (which usually represents the whole page). Genre and design patterns have a social background. On this background, there are different groups of people with the same interests.

Definition 13.1. (Web) microgenre is a part of a Web page,

1. Whose purpose is general and repeats frequently
2. Which can be named intelligibly and more or less unambiguously so that the name is understandable for the Web page user (developer, designer, etc.)
3. Which is detectable on a Web page using computer algorithm

The microgenre can, but does not have to, strictly relate to the structure of a Web page in a technical sense, e.g. it does not necessarily have to apply that it is represented by one subtree in the DOM tree of the page or by one block in the sense of the visual layout of the page. Rather it can be represented by one or more segments of a page, which form it together (see Fig. 13.4).

Remark 13.1. Microgenres are also contexts which encapsulate related information. In paper [14] we show the way we extract snippets from individual microgenres.

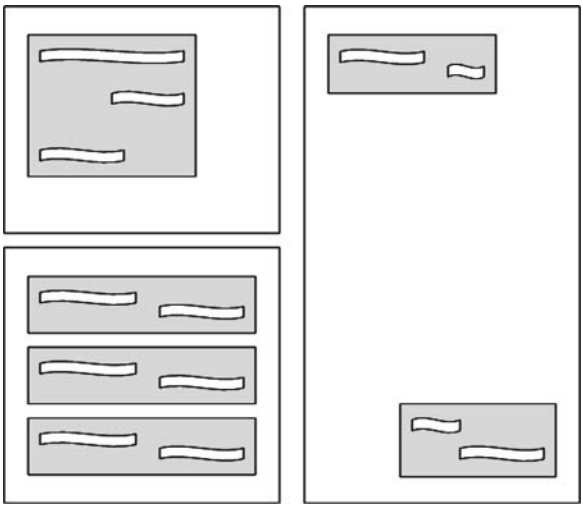


Fig. 13.4 Microgenres formed by Web page segments

We use these snippets in our Web application as an additional information for Web page description. The detection of microgenres can be considered in a similar way as the first step for using Web information extraction methods (see [8]).

13.2.1.1 Microgenre Recognition

It follows the previous description that in order to be able to speak about the microgenre, this element has to be distinguishable by the user. From what attributes should the user recognize, if and what the Microgenre is in question? We work with up to three levels of view:

1. The first view is purely semantic in the sense of the textual content of a page. It does not always need to have to be a meaning in a sense of natural language such as sentences or paragraphs with a meaningful content. Logically coherent data blocks can still lack in grammar (see [35]).

   For example, price information can be only a group of words and symbols ('price', 'vat', symbol $) of a data type (price, number). For similar approach see [27].
2. The second view is visual in the sense of page perception as a whole. Here individual segments of perception or groups of segments of the page are in question. It is dependent on the use of colors, font and auxiliary elements (lines, horizontal and vertical gaps between the segments, etc.) Approaches based on visual analysis of Web pages can be found in [6, 30].

3. The third view is a structural one in a technical sense. It is about the use of special structures, such as tables, links, navigation trees, etc. There are approaches based on the analysis of the DOM tree and special structures as tables [20, 25].

The first view is dependent on the user's understanding of the text stated on a Web page. The second and third views are independent of this user ability. However, it can be expected that an Arabic or Chinese product page will be recognized also by an English-speaking user who does not have a command of those languages. It is determined by the fact that for the implementation of certain intentions there are habitual procedures which provide very similar results regardless of the language. On the other hand, if the user understands the page, he or she can focus more on the semantic content of the microgenre. For example, in the case of "product info", the user can read what the product in question is, what its price is and on what conditions it can be purchased.

13.2.2 Web Page Description

Using mentioned views, we can describe every Web page by genre or a group of microgenres. This description, in principle, defines the communities mentioned in the preface of our chapter. On the other side, each defined community contributes somehow to the development of the Web and simultaneously to the behavior of the related communities (e.g., communities involved in blogging as mentioned earlier). Consequently, these contemplations lead to reduction of the Web to some particular types of pages and communities. This can be very useful, for example, in searching. Knowing that the user prefers product pages (or review pages, discussion pages, etc.), we can help him or her on – namely on the basis of knowledge – how the community of developers in a given domain (Web design patterns) works and what is the prevailing intent of the pages (genres).

13.3 Related Work

When we perceive the Web page as whole, the purpose is represented by a so-called Web genre. Similarly, the view of individual segments of the Web page is closely related to Web design patterns. A Web genre is a taxonomy that incorporates the style, form and content of a document which is orthogonal to the topic, with fuzzy classification to multiple Web genres [4]. For classification, there are many approaches and also many methods for genre identification. Kennedy and Shepherd [12] analyzed home page genres (personal home page, corporate home page or organization home page). Chaker and Habib [7] proposed a flexible approach for Web page genre categorization. Flexibility means that the approach assigns a document to all predefined genres with different weights. Dong et al. [10] described a set of experiments to

examine the effect of various attributes of Web genre for the automatic identification of the genre of Web pages. Four different genres were used in the data set (FAQ, News, E-Shopping and Personal Home Pages).

Rosso [26] explored the use of genre as a document descriptor in order to improve the effectiveness of Web searching. Author formulated three hypotheses: (1) Users of the system must possess sufficient knowledge of the genre. (2) Searchers must be able to relate the genres to their information needs. (3) Genre must be predictable by a machine-applied algorithm because it is not typically explicitly contained in the document.

Design patterns and pattern languages came from the architecture in the work of Christopher Alexander et al. [2]. From the mid sixties to mid seventies, Alexander et al. defined a new approach to architectural design. The new approach centered on the concept of pattern languages is described in a series of books [2]. Alexander's definition of pattern is as follows: "Each pattern describes a problem, which occurs over and over again in our environment, and then describes the core of the solution to that problem, in such a way that you can use this solution a million times over, without ever doing it the same way twice."

According to Tidwell [31], patterns are structural and behavioral features that improve the applicability of software architecture, a user interface, a Web site or something or another in some domain. They make things more usable and easier to understand. Patterns are descriptions of best practices within a given design domain. They capture common and widely accepted solutions, and their validity is empirically proved. Patterns are not novel, patterns are captured experiences and each of their implementation is a little different.

Good examples are also the "Web design patterns", which are patterns for design related to the web. A typical example of a Web design pattern can be the forum pattern (see Fig. 13.5). This pattern is meant for designers who need to implement this element on an independent Web page or as a part of another Web page. The pattern describes key solution features without implementation details.

Fig. 13.5 Sample of Forum pattern (www.welie.com)

Generally, the design patterns describe a proven experience of repeated problem solving in the area of software solution design. From this point of view, the design patterns belong to key artifacts securing efficient reuse. While the design patterns have been proven in real projects, their usage increases the solution quality and reduces the time of their implementation.

There is a wide area of methods, which aim to detect objects related to patterns and extract their semantic details (e.g., opinion extraction [16], news extraction [34], Web discussion extraction [19], product detail extraction [24], and technical features extraction [28]).

13.4 Tools and Techniques

In the beginning of this chapter, we have presented our motivation, goals and thoughts. Now we mention some basic notions of useful technologies. The subsequent experiments illustrate the concrete application of proposed approach.

13.4.1 Pattrio Method

In our approach, we were inspired by the design pattern use for the analysis of Web page content. If we look for microgenres on Web pages, we need detailed technical information. That is why we have created our own catalog, in which we describe those repeated microgenres, which we manage to detect on Web pages by our method. For description we use a description similar to a pattern description, but its intention is different and it aims at understanding what characteristics are important for detection algorithm design. However, our view has a lot in common with patterns. It is mainly because also for us, in the same way as for a pattern, the most important characteristic is the name of the thing described. Our approach is different on the level of the general view and target. Simply said, we understand microgenre used by us as a projection of a Web design pattern . This projection does not always have to be unambiguous, e.g. one pattern can be projected to more microgenres.

13.4.1.1 Pattrio Catalog

Patterns are designed for Web designers who work with them and use them in production. A pattern description is composed from parts and each part describes a specific pattern feature. Authors usually use the pattern structure introduced in [2]. In the description, there is a pattern name, problem description, context, solution and examples of use. Usually, these are also consequences of the use of the pattern and related patterns which relate somehow with the pattern being used. For our description of microgenre we use the similar section-oriented structure.

13.4.1.2 Example – Discussion (Forum)

Problem How can a discussion about a certain topic be held? How can a summary of comments and opinions be displayed?

Context Social field, community sites, blogs, etc. Discussions about products and service sales. Review discussions. News story discussion.

Forces A page fragment with a headline and repeating segments containing individual comments. Keywords to labeling discussion on the page (discussion, forum, re, author, ...). Keywords to labeling persons (first names, nicknames). Date and time. There may be a form to enter a new comment. Segments with the discussion contributions are similar to the mentioned elements view, in form.

Solution Usually, an implementation using a table layout with an indentation for replies (or similar technology leading to the same-looking result) is used. The discussion is often together with the login. If discussion is on selling a product on Web site, there are usually purchase possibility, price information. The discussion can be alone on the page. In other case, there is also the something to read. In different domains the discussion can be displayed with review, news, etc. See Fig. 13.6.

Fig. 13.6 Discussion

13.4.1.3 Detection Algorithm

We have defined sets of elements mentioned above for each microgenre that are characteristic for this microgenre (words, data types, technical elements). These elements have been obtained on the basis of deeper analysis of a high volume of Web pages. This analysis also included the calculation of the weight of individual microgenre elements that defines the level of relevance for the microgenre. Besides, we have implemented a set of partial algorithms whose results are the extracted data types and also the score for the quality of fulfillment of individual rules of microgenres.

In 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 microgenre instance can be found (technical elements). The rules are the way that individual elements take part in the microgenre display. While defining these rules, we have been inspired by the Gestalt principles (see Fig. 13.7 and [31]). We formulated 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 to create of independent closed segments containing the elements. One or more segments then create the microgenre instance on the Web page. The third one (similarity) defines that the microgenre includes more related similar segments. The forth one (continuity) defines that the microgenre contains various segments that together create the Web pattern instance. The relations among microgenres can be on various levels similar as classes in OOP (especially simple association and aggregation).

The basic algorithm for detection of microgenres 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., see Table 13.1), segmentation and evaluation of rules and associations. The result for the page is the score of

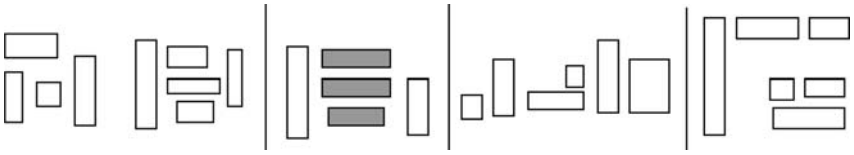


Fig. 13.7 Gestalt principles (proximity, similarity, continuity, closer)

Table 13.1 HTML tags – classification for analysis

Types	Tags
Headings	H1, H2, H3, H4, H5, H6
Text containers	P, PRE, BLOCKQUOTE, ADDRESS
Lists	UL, OL, LI, DL, DIR, MENU
Blocks	DIV, CENTER, FORM, HR, TABLE, BR
Tables	TR, TD, TH, DD, DT
Markups	A, IMG
Forms	LABEL, INPUT, OPTION

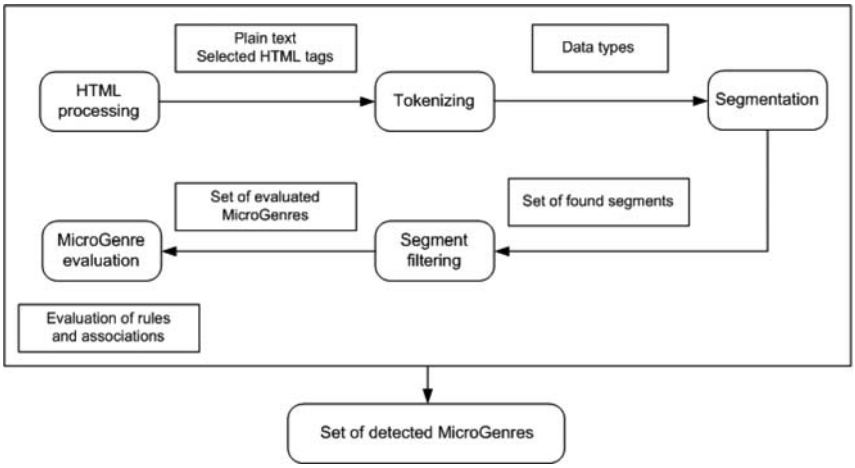
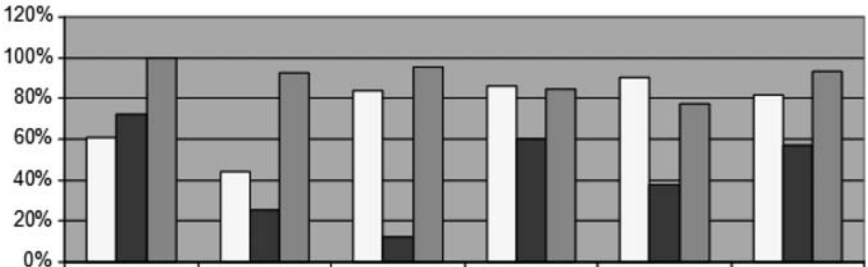


Fig. 13.8 Microgenre detection process

Algorithm 13.1 Microgenres score (pseudocode)

```
input  : Set of PageEntities, set of microgenres
output : MicroGenresScore
foreach PageEntity in PageEntities do
    if PageEntity is MicroGenreEntity then
        if does not exist segment then
            create new segment in list of segment;
            to add page entity to segment;
        end
        add page entity to segment;
    end
end
foreach segment in list of segments do
    compute proximity of segment;
    compute closure of segment;
    compute Score(proximity, closure) of segment;
    if Score is not good enough then
        remove segment from list of segments;
    end
end
compute similarity of list of segments;
compute continuity of list of segments;
compute Score(similarity, continuity) of microgenre;
return Score
```

microgenres that are present on the page. The score then says what is the probability of expecting the microgenre instance on the page for the user. The entire process, including microgenre detection, is displayed in Fig. 13.8 and Algorithm 13.1.



	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. 13.9 Accuracy of Pattrio method

13.4.1.4 Method Accuracy

The accuracy of the proposed method is about 80% (see [13]). Figure 13.9 shows the accuracy of the Pattrio method for three selected products (Apple iPod Nano 1 GB, Canon EOS 20D, Star Wars Trilogy film) and for the *discussion* and the *purchase possibility* microgenres. We used only the first 100 pages for each product. We manually, and using Pattrio method, evaluated the pages using a three-degree scale [9]:

+ Page does not contain required microgenre.
? Unable to evaluate results.
− Page do not contain required microgenre.

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

13.4.2 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 the presence of a microgenre in the page content. We are able to describe the amount of its presence at some ref:CSNA-13-09le 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 [9] and a very imprecise environment, we should consider several practical issues, which have to be solved prior to the first application. Methods of matrix decomposition have succeeded in reducing the dimensions of input data (see [29] for application connected with formal concept analysis and [17, 18] for overview).

13.4.2.1 FCA Basics

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 is called **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 the relation between G and M. The elements of G are defined as objects and the elements of M as attributes of the context. In order to express that an object $g \in G$ is related to I with the attribute $m \in M$, we record it as gIm or $(g, m) \in I$ and read that the object g has the attribute m.

For a set $A \subseteq G$ of objects we define $A^{'} = \{m \in M \mid gIm \; for \; all \; g \in A\}$ (the set of attributes, common to the objects in A). Correspondingly, for a set $B \subseteq M$ of attributes we define $B^{'} = \{g \in G \mid gIm \; for \; all \; m \in B\}$ (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 $B^{'} = A$. We call A the extent and B the intent of the concept $(A, B) \cdot \mathcal{B}(G, M, I)$ denotes the set of all concepts of context (G, M, I) and forms a complete lattice (so called **Gallois lattice**). For more details, see [11].

Now we give only a brief overview of FCA in fuzzy environment (approach of Bělohlávek et al.). Instead of classical binary case, we can consider the so-called complete residuated lattice $\mathbf{L} = \langle L, \vee, \wedge, \otimes, \rightarrow, 0, 1\rangle$, where $\langle L, \vee, \wedge, 0, 1\rangle$ is a complete lattice (with 0 and 1 being the smallest and biggest element), $\langle L, \otimes, 1\rangle$ is a commutative monoid and $\langle \otimes, \rightarrow\rangle$ is an adjoint pair of binary operations (truth functions of fuzzy conjunction and fuzzy implication). The notion of being the element of a set can be replaced by the degree in which the element is contained in the set $(A(x))$. The notion of subset can be inferred in a similar way.

Now we can define the fuzzy **L-context** as $\mathbf{L} = \langle X, Y, I\rangle$ with X as a set of objects, Y as a set of attributes and I as a fuzzy relation $I : X \times Y \rightarrow L$. For a fuzzy set $A \in \mathbf{L}^X$, $B \in \mathbf{L}^Y$ (A is a fuzzy set of objects, B is a fuzzy set of attributes), we define fuzzy set of attributes $A^{'} \in \mathbf{L}^Y$ and fuzzy set of objects $B^{'} \in \mathbf{L}^X$ as

$$A^{'}(y) = \bigwedge_{x \in X} \left(A(x)^{*_X} \rightarrow I(x, y)\right),$$
$$B^{'}(x) = \bigwedge_{y \in Y} \left(B(y)^{*_Y} \rightarrow I(x, y)\right).$$

The ^{*_X} and ^{*_X} are the so-called truth-stressing functions or simple hedges which allows us to control the size of the resulting lattice.

Formal fuzzy concept is a pair $\langle A, B\rangle$, $A \in \mathbf{L}^X$, $B \in \mathbf{L}^Y$, such that $A^{'} = B$ and $B^{'} = A$. In this case, we will call A as the extent of the concept, and B as its intent. As $\langle \mathcal{B}(X^{*_X}, Y^{*_Y}, I), \leq\rangle$, we denote the set of all concepts, which when accompanied by the induced order is called **fuzzy concept lattice**. For further details and comparison with other approaches please see [3]. As you can see, the key terms from classical FCA have their parallel in the fuzzy environment. So the key algorithms have.

13.5 Experiments

In this chapter, 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 microgenres as the basis of this architecture, because they encapsulate those parts of the page content that have partial intent. We use our Pattrio method for microgenre 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 can 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 a special 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 us take an e-shop as an example. It is sure that the aim of the e-shop owner is to sell the most 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, he or she 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 or she 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 can be seen in the development of Web pages with usual intent. This unification is based on principles which come 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 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.

We 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). Users from a group of students and teachers of high schools and VŠB – Technical University Ostrava, Czech Republic – were using this application for more than one year to search for everyday information. We have not influenced the process of searching in any way. The purpose of this part of the experiment was to view the World Wide Web using the perspective of users (as the search engines play key role in World Wide Web navigation). In the end, we obtained data set with more than 115,000 Web pages. After cleaning up, 77,850 unique Czech pages remained. For every single Web page we have performed the detection of 16 microgenres. The page did not have to contain any microgenre, as well as it may have theoretically contained 16 microgenres (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). The names of microgenres emerged from the discussions between us and students that took part in our experiments. They are therefore outcomes of social interaction. We used such preprocessed data sets for all the experiments.

13.5.1 Web Site Visualization

For the visualization of extracted information we have adopted one common graph drawing method, which works as follows: in the center of Fig. 13.10 you can see the circle representing the Web site. The size of the circle is determined by the relative size (number of pages) in the data set. The circles around the Web site correspond to individual microgenres. The size of circles is again determined by their relative presence in the data set.

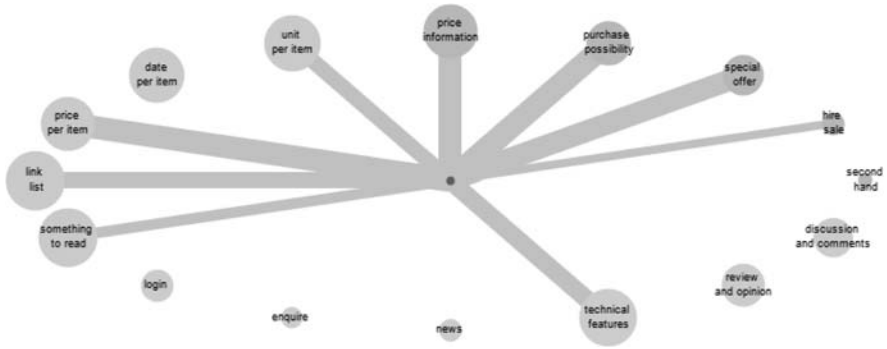


Fig. 13.10 Typical Web site aimed at selling products

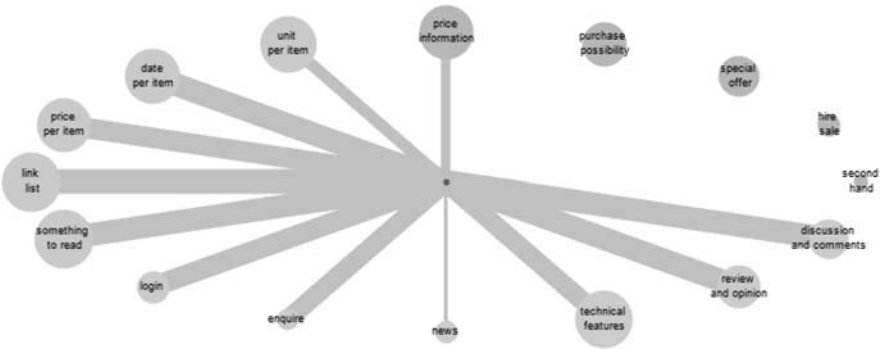


Fig. 13.11 Typical Web site aimed at information sharing

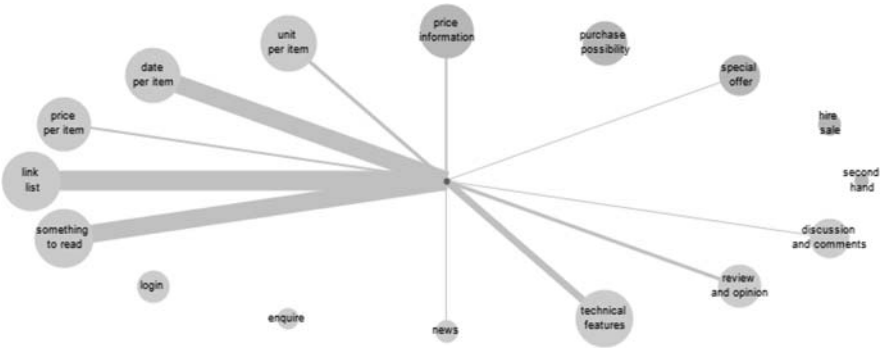


Fig. 13.12 Typical university Web site

The Web site is connected to microgenres using a straight line. The strength of this connection (represented by the line width) corresponds to the detected degree of microgenres present in the Web site.

Figures 13.10, 13.11, and 13.12 contain the described visualization of some Web sites from the Czech Republic – typical Web sites aimed at selling products, sharing information and education.

Figure 13.13 depicts two Web sites with similar intent – selling products. The second one differs in allowing users to share information in addition to a purchase possibility.

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

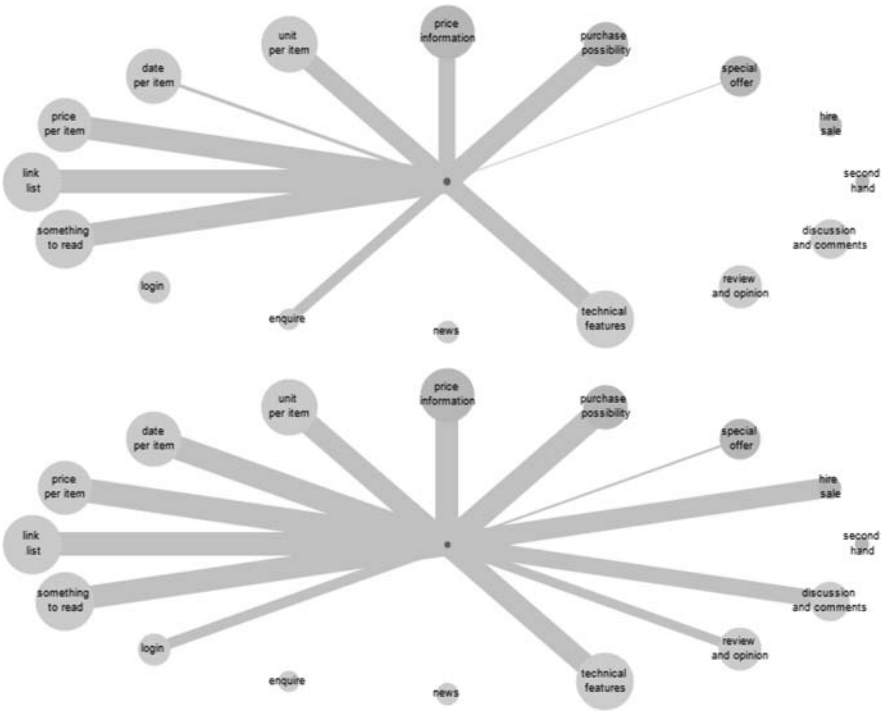


Fig. 13.13 Comparison of two product-oriented Web sites

13.5.2 Web Site Clustering

In the next 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 data set have been taken into consideration. Each domain is accompanied by detected microgenres. This list is transformed into a binary matrix and considered as a formal context. Using methods of FCA we have computed a conceptual lattice which can be seen in Fig. 13.14. The resulting matrix has 516 rows (objects) and 16 columns (attributes) and the computed conceptual lattice contains 378 concepts.

From the computed lattice we have selected a sub-lattice containing 18 Web sites dealing with cell phones. Only five attributes have been selected and the visualization was created in a slightly different manner (see Fig. 13.15 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 the whole set of Web sites can be divided into two groups – the first one

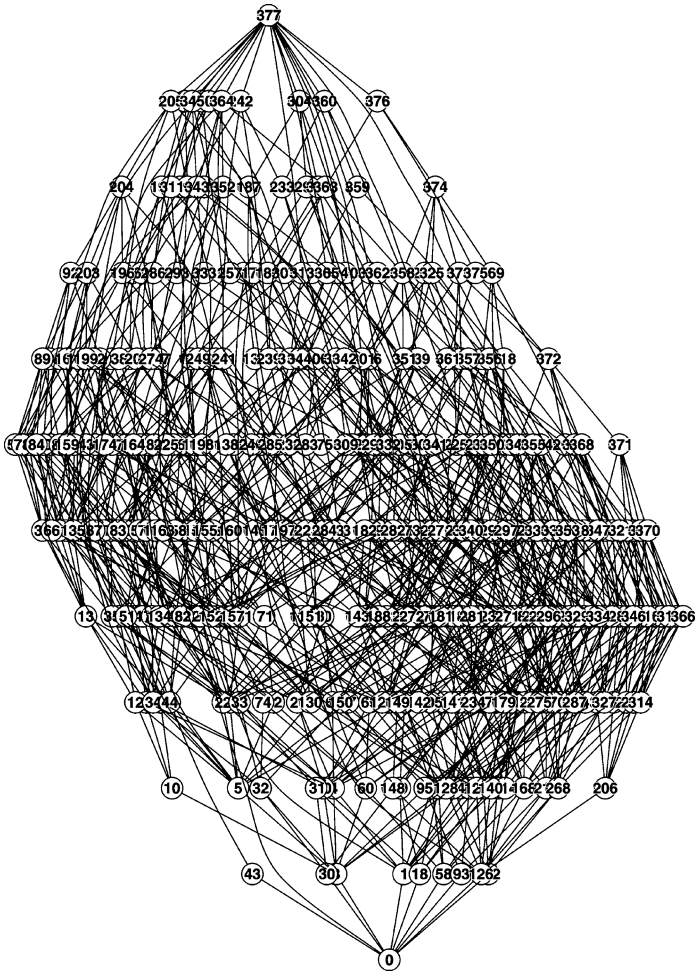


Fig. 13.14 Lattice calculated from whole dataset

contains sites where users are enabled to buy cell phones and the second one where the users are allowed to have a discussion. Deeper insight gives you more detailed information about Web site structures and relations.

The conceptual lattice forms a graph, which can be interpreted as an expression of relation between different Web sites. As a result, it describes the relation between different Web communities.

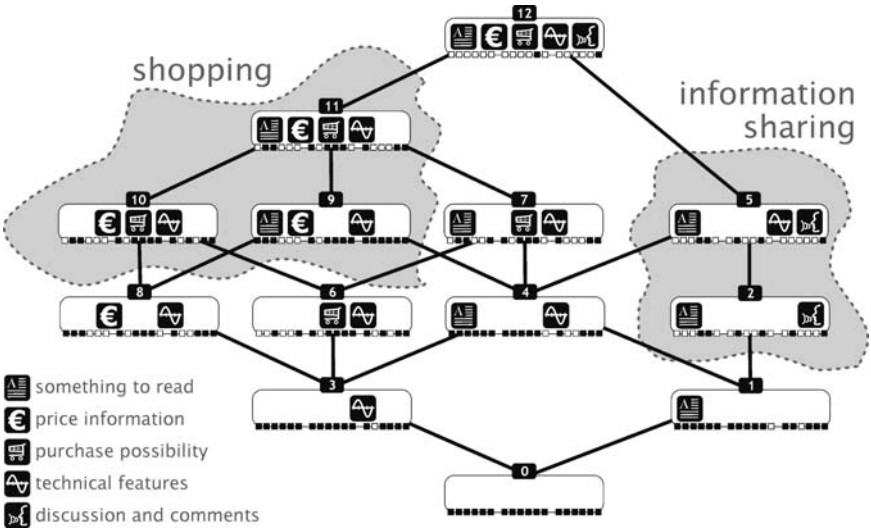


Fig. 13.15 Part of lattice

13.6 Conclusions and Future Work

In this chapter, 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. 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 if we focus on Web sites and the Web page content they provide, we might come across a variety of interesting questions. These questions may bear upon the Web sites' similarity and the similarity of social groups involved in these pages, which could formulate interesting future research directions. The identification of additional microgenres and design of specific algorithms for their detection will be the matter of our future research.

References

1. Adamic L, Adar E (2005) How to search a social network. J Soc Networks 27:187–203
2. Alexander Ch (1977) A pattern language: towns, buildings, construction. Oxford University Press, New York
3. Belohlavek R, Vychodil V (2005) What is a fuzzy concept lattice. In: Proceedings of the CLA, 3rd international conference on concept lattices and their applications, Olomouc, Czech Republic, pp 34–45

4. Boese ES (2005) Stereotyping the web: genre classification of Web documents. Colorado State University
5. Borchers JO (2000) Interaction design patterns: twelve theses, workshop, vol 2. The Hague, The Netherlands
6. Cai D, Yu S, Wen JR, Ma WY (2003) Extracting content structure for Web pages based on visual representation. In: Fifth Asia Pacific Web conference, Xian, China, pp 406–417
7. Chaker J, Habib O (2007) Genre categorization of Web pages. In: Proceedings of the 7th IEEE international conference on data mining workshops, Omaha, Nebraska, USA, pp 455–464
8. Chang H Ch, Kayed M, Girgis MR, Shaalan KF (2006) A survey of Web information extraction systems. IEEE Trans Knowl Data Eng 18:1411–1428
9. Cole RJ, Eklund PW (1999) Scalability in formal concept analysis. Comput Intell 15:11–27
10. Dong L, Watters C, Duffy J, Shepherd M (2008) An examination of genre attributes for Web page classification. In: Proceedings of the 41st annual Hawaii international conference on system sciences, Big Island, HI, USA, pp 133–143
11. Ganter B, Wille R (1997) Formal concept analysis: mathematical foundations. Springer, New York
12. Kennedy A, Shepherd M (2005) Automatic identification of home pages on the Web. In: Proceedings of the 38th Hawaii international conference on system sciences, Big Island, HI, USA
13. Kocibova J, Klos K, Lehecka O, Kudelka M, Snasel V (2007) Web page analysis: experiments based on discussion and purchase Web patterns. In: Web intelligence and intelligent agent technology workshops, Silicon Valley, CA, USA, pp 221–225
14. Kudelka M, Snasel V, Lehecka O, El-Qawasmeh E, Pokorny J (2008) Web pages reordering and clustering based on Web patterns, SOFSEM 2008: Conference on current trends in theory and practice of computer science, Novy Smokovec, Slovakia, pp 731–742
15. Kumar R, Novak J, Tomkins A (2006) Structure and evolution of online social networks. In: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, Philadelphia, PA, USA, pp 611–617
16. Lee D, Jeong OR, Lee S (2008) Opinion mining of customer feedback data on the web. In: Proceedings of the 2nd international conference on Ubiquitous information management and communication, Suwon, Korea, pp 230–235
17. Lee D, Seung H (1999) Learning the parts of objects by non-negative matrix factorization. Nature 401:788–791
18. Letsche T, Berry MW, Dumais ST (1995) Computational methods for intelligent information access. In: Proceedings of the 1995 ACM/IEEE supercomputing conference, San Diego, CA, USA
19. Limanto HY, Giang NN, Trung VT, Zhang J, He Q, Huy NQ (2005) An information extraction engine for web discussion forums. In: International World Wide Web conference, Chiba, Japan, pp 978–979
20. Liu B, Grossman R, Zhai Y (2003) Mining data records in Web pages. KDD 2003, ACM SIGKDD international conference on knowledge discovery and data mining, Washington, DC, USA, pp 601–606
21. Martin JR (1995) Text and clause: fractal resonance. Text 15:5–42
22. Murata T (2004) Discovery of user communities from Web audience measurement data. Web Intelligence 2004, pp 673–676
23. Murata T, Takeichi K (2007) Discovering and visualizing network communities. Web intelligence/IAT workshops 2007, pp 217–220
24. Nie Z, Wen JR, Ma WY (2007) Object-level vertical search. In: Third biennial conference on innovative data systems research, Asilomar, CA, USA, pp 235–246
25. Pivk A, Cimiano P, Sure Y, Gams M, Rajkovic V, Studer R (2007) Transforming arbitrary tables into logical form with TARTAR. Data Knowl Eng 60:567–595
26. Rosso MA (2008) User-based identification of Web genres. JASIST (JASIS) 59(7):1053–1072
27. Santini M (2009) Description of 3 feature sets for automatic identification of genres in web pages. www.nltg.brighton.ac.uk/home/Marina.Santini/three_feature_sets.pdf. Accessed on 30 April 2009

28. Schmidt S, Stoyan H (2005) Web-based extraction of technical features of products. Beiträge der 35. Jahrestagung der Gesellschaft für Informatik, pp 256–261
29. Snasel V, Polovincak M, Dahwa HM, Horak Z (2008) On concept lattices and implication bases from reduced contexts. In: Supplementary proceedings of the 16th international conference on conceptual structures. ICCS 2008, pp 83–90
30. Takama Y, Mitsuhashi N (2005) Visual similarity comparison for Web page retrieval. In: Web intelligence, Compiegne, France, pp 301–304
31. Tidwell J (2005) Designing interfaces: patterns for effective interaction design. O'Reilly, Sebastopol, CA, USA, pp 0–596
32. Toyoda M, Kitsuregawa M (2001) Creating a Web community chart for navigating related communities. Hypertext 2001, Aarhus, Denmark, pp 103–112
33. Van Duyne DK, Landay JA, Hong JI (2003) The design of sites: patterns, principles, and processes for crafting a customer-centered Web experience. Addison-Wesley Professional, USA
34. Zheng S, Song R, Wen JR (2007) Template-independent news extraction based on visual consistency. In: Proceedings of the 22nd AAAI conference on artificial intelligence, Vancouver, British Columbia, Canada, pp 1507–1513
35. Zhu J, Zhang B, Nie Z, Wen JR, Hon HW (2007) Webpage understanding: an integrated approach. In: Conference on knowledge discovery in data, San Jose, California, USA, pp 903–912