Journal of the Applied Mathematics, Statistics and Informatics (JAMSI), 3 (2007), No. 1

Web Pattern Oriented Semantics of Web Pages[1]

VÁCLAV SNÁŠEL, MILOŠ KUDĚLKA, ONDŘEJ LEHEČKA

Abstract

In this paper we introduce a new approach to semantic analysis of web pages. To prove this approach we designed method for analysis and evaluation of web pages. The method is built on a silent agreement between web designers and users. The key aspects of this agreement are web patterns which are used by web designers in their web page implementations. With our method we can find out whether the pattern is presented on the page with high level of relevance. Key feature of our method is independence on the page HTML code. In this paper we explain essentials of our approach as well as key features of our method and context for proper usage. We present experiments which prove efficiency of the method and usefulness of the approach.

Mathematics Subject Classification 2000: 68T20
General Terms: Experimentation, Human Factors
Additional Key Words and Phrases: Semantic Analysis, Web Pattern, Human-Computer Interaction

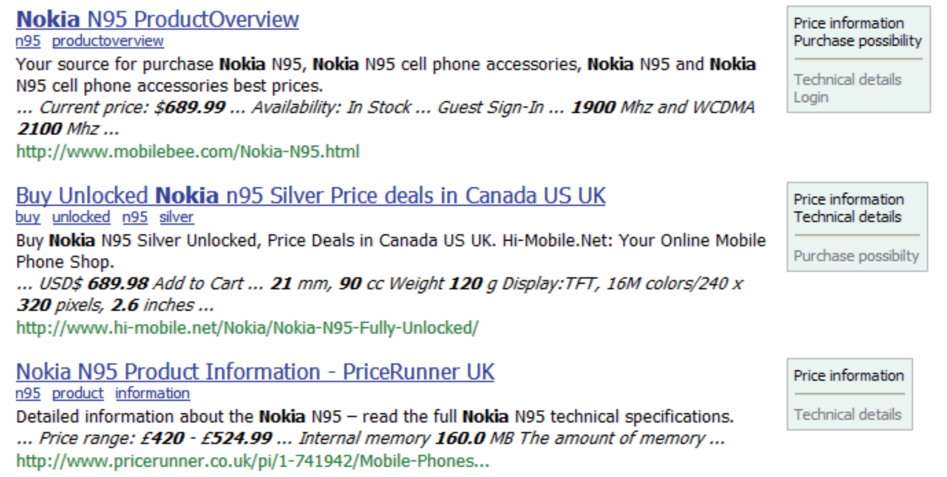
1. INTRODUCTION

One way to help the user in orientation within a vast amount of non-structured data is clustering according to common key properties. The biggest problem, however, remains in the definition what is the key property which is useful for definition of similarity [10]. In our approach we work with web pages and web patterns which are presented on these web pages. The web patterns provide, on a certain level, a formalized mechanism for the description of common features of an object which is commonly visible on web pages. For this purpose we developed some new web patterns, extended web pattern description and created their taxonomy.

The obvious assumption of web page availability is its presence in search engines indexes. The presence of the web page in the index does not mean that the page is always available through simple query. Different methods are used for measuring the relevance of a web page against the query. Considering the huge amount of web pages it is noticeable that current methods are not sufficient. One important feature of our approach is that it brings new information about web pages which is not currently used but the information is readable for users (users understand it). Using this information it is possible to provide additional criterions for measuring relevance of the web page against the user's expectance. Our tests and experiments with users (see [9]) prove the presented

[1] This research was supported by the Funding Agency.

approach can help with orientation in a vast amount of non-structured data. If we can reveal which web patterns a web page contains user can immediately create a notion about what kind of information can expect (see Fig. 1).



Nokia N95 ProductOverview
n95 productoverview
Your source for purchase **Nokia** N95, **Nokia** N95 cell phone accessories, **Nokia** N95 and **Nokia** N95 cell phone accessories best prices.
... Current price: ***$689.99*** *... Availability: In Stock ... Guest Sign-In ...* ***1900*** *Mhz and WCDMA* ***2100*** *Mhz ...*
http://www.mobilebee.com/Nokia-N95.html

Price information
Purchase possibility
Technical details
Login

Buy Unlocked **Nokia** n95 Silver Price deals in Canada US UK
buy unlocked n95 silver
Buy **Nokia** N95 Silver Unlocked, Price Deals in Canada US UK. Hi-Mobile.Net: Your Online Mobile Phone Shop.
... USD$ ***689.98*** *Add to Cart ...* ***21*** *mm,* ***90*** *cc Weight* ***120*** *g Display:TFT, 16M colors/240 x* ***320*** *pixels,* ***2.6*** *inches ...*
http://www.hi-mobile.net/Nokia/Nokia-N95-Fully-Unlocked/

Price information
Technical details
Purchase possibilty

Nokia N95 Product Information - PriceRunner UK
n95 product information
Detailed information about the **Nokia** N95 – read the full **Nokia** N95 technical specifications.
... Price range: ***£420*** *-* ***£524.99*** *... Internal memory* ***160.0*** *MB The amount of memory ...*
http://www.pricerunner.co.uk/pi/1-741942/Mobile-Phones...

Price information
Technical details

Fig. 1. Query from selling product domain.

On Figure 1 there is a sample of three searched pages on the query "*nokia*" in our experimental search engine *www.pattrio.net*. There is a "tie-on label" on the right side of each web page. The label contains a list of automatically detected web patterns (font weight was used to designate the relevance of the detection; bold font is designating high relevance). Web page snippets contain best segments from the web patterns found on the page. The segments are highlighted in italics.

Web patterns are detected using analysis and segmentation of HTML and text content [9]. Whole process is in Fig. 2.

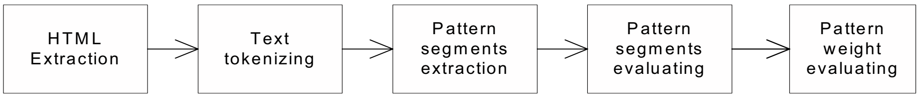

HTML Extraction
Text tokenizing
Pattern segments extraction
Pattern segments evaluating
Pattern weight evaluating

Fig. 2. Web pattern detection process.

The rest of the paper is organized as follows. Section 2 presents related work. Section 3 presents the web patterns basics. Section 4 presents samples of web patterns from our catalog. In section 5 we describe page similarity based on web patterns. Section

6 and 7 contains a description of experiments and analysis of results, and finally section 8 is the conclusion.

2. RELATED WORK

In the field of non-structured data analysis there are various approaches (for example information retrieval, semantic web, information extraction...). The construction of web information extraction systems is an approach similar to ours. There web pages are transformed into program-friendly structures such as a relational database. In this approach they need to analyze the structure and templates of the web page. The survey of major web data extraction approaches is presented in [3]. An interesting approach is mentioned in [10]. By empirically studying web pages across web sites about the same type of objects, they found out many HTML template independent features. In [11] is presented a method for extracting objects from the web. Authors have written that the main challenge for extraction is that objects of the same type are distributed among diverse web sources, whose structures are highly heterogeneous. In this point our method is related to theirs. The next approach related to ours is in [15]. They propose a vision-based page segmentation algorithm to detect the semantic content structure in a web page. The next similar approach is mentioned in [7]. It analyzes chosen parts of pages to obtain structured domain specific information (tourism domain). Other similar approaches are automatic transformation of arbitrary table-like structures into knowledge models [12] or domain-oriented approach to web data extraction based on a tree structure analysis [13]. There is an interesting conjunction with a paper [6]. Authors of the paper analyzed web pages focusing on web site patterns. In three different time intervals they observed how web designers have changed web design practices. They also realized that the content of web pages remains the same whereas a form is being developed so it better fulfils users' expectation. Our work confirms results mentioned in the paper. For us such characteristics of web patterns that are independent of the web page design are very important.

3. WEB PATTERN

In this section, we first introduce the concept of Web patterns and pattern extraction. By the term "web pattern" we mean any high-level object which is on a certain (relatively high) level of abstraction presented often and repeatedly on web pages.

We especially consider the web pattern only such an object which can be named so it is clear from the name what the pattern describes (both users and developers should agree on this). Examples of such web patterns are the *Price information* pattern or the *Purchase possibility* pattern (you will find the description below). These web patterns come along with pages about selling products or services. Both mentioned web patterns provide semantic information to the user. This is one of the key features. There are also web patterns which do not contain any semantic content. An example of such a web pattern can be the *Something to read* pattern describing pages which contain a bigger amount of text. Another example is the *Link list* pattern which describes pages containing a group of links. You will find description of those web patterns below.

It is very important for the next reading that for the purpose of this paper we work only with those web patterns which can be automatically found on a web page based on their description with a relatively high relevance. For example, the *Welcome page* can also be considered a web pattern. But the detection of this pattern would be much harder (or even impossible due to the designer creativity).

3.1. Web Pattern Taxonomy

We divide web patterns into two groups (many details about patterns and pattern classification are in [4]). In first group there are web patterns providing semantic information to the user (*Domain patterns*). This semantic information is connected to a certain domain which the page and expected users belong to. The domain is considered to be composed by

1. web pages with specific content
2. users with specific requirements and expectations.

An example of domains can be the selling product and services domain, tourism domain, culture domain, newscast domain, etc. The advantage of mentioned domains is that there are a lot of patterns described in catalogs [5][14]. Some web patterns which we use in our approach are adopted from the catalogs.

The second group contains web patterns independent of a domain. They describe more structural and technical features of typical solutions (*Structural patterns*).

3.2. Pattern Description and Structure

Patterns are designated for users (web designers in this case) who work with them and use them in production. A pattern description is composed from parts and each part is describing specific pattern feature. Authors usually use the structure introduced in [1]. 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 we use the structure originated by Kent Beck [16]. There is also a *Forces* section describing details which can help in the automatic detection of web pattern on a web page. The description of such details comes out from our experiments with web pattern detection in a vast amount of web pages. The description also helps us to understand how to design detecting algorithms. The example descriptions and examples themselves presented in a *Solution* section show how different designers can proceed in the implementation of pages. In further text we will describe some web patterns.

Title – appropriate pattern name
Problem: A single brief sentence describing the problem which pattern solves.
Context: A list of situations where the pattern occurs.
Forces: A list of details which influence the pattern identification. We are focusing especially on features useful for automatic detection.
Solution: Description of the solution with examples.

The related patterns are also very important. If there is such a pattern in the description, we highlight the name with italic font.

4. WEB PATTERN SAMPLES

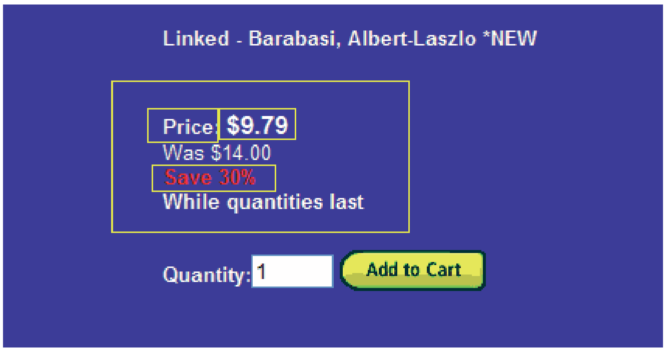
We choose patterns from a collection (or corpus) **C** which we use for automatic detection on web pages. We use more than twenty web patterns (domain and structural) for the analysis. We choose three domain and two structural patterns.

4.1. Domain Pattern - Price Information

Problem: How to graphically show selling information to the user?
Context: Selling products, services, etc.
Forces: A page fragment alone usually bound on a small space. Keywords labeling that there is a price. Numbers in combination with currency sign. There is usually a picture

and product identification near each other on the page. All mentioned elements are on the small space for a current product.

Solution: In different contexts there should be more different implementations. It is the case of pages with a single offer or pages with more offers (catalog with offers). The patterns may occur at a page border as an advert. The pattern is usually present together with patterns *Purchase possibility*, *Special offer* and *Repayment*. See Fig 3.



Linked - Barabasi, Albert-Laszlo *NEW

Price: $9.79
Was $14.00
Save 30%
While quantities last

Quantity: 1

Fig. 3. *Price information* pattern.

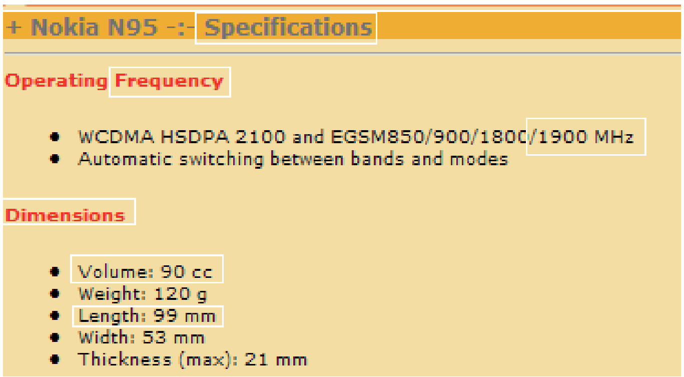
4.2. Domain Pattern - Technical Details

Problem: How to lucidly show technical information of a product?

Context: Selling products like electronics, appliances (fridge, etc.). Personal website (for example a product fan). Manufacturer's website.

Forces: A page fragment with a headline and a list of single rows describing product parameters. Key words labeling details section on the page (details, parameters…). Keywords labeling parameters (size, weight, frequency…). Numbers in combination with unit sign (cm, kg, MHz…). All mentioned pattern elements are placed on bigger space of page so the user can continuously read them.

Solution: Usually an implementation using a table layout (or similar technology leading to the same-looking result) is used. If the pattern is on a selling product website there are usually *Purchase possibility*, *Selling information* patterns and often *Login* pattern. See Fig. 4.



+ Nokia N95 -:- Specifications

Operating Frequency

- WCDMA HSDPA 2100 and EGSM850/900/1800/1900 MHz
- Automatic switching between bands and modes

Dimensions

- Volume: 90 cc
- Weight: 120 g
- Length: 99 mm
- Width: 53 mm
- Thickness (max): 21 mm

Fig. 4. *Technical details* pattern.

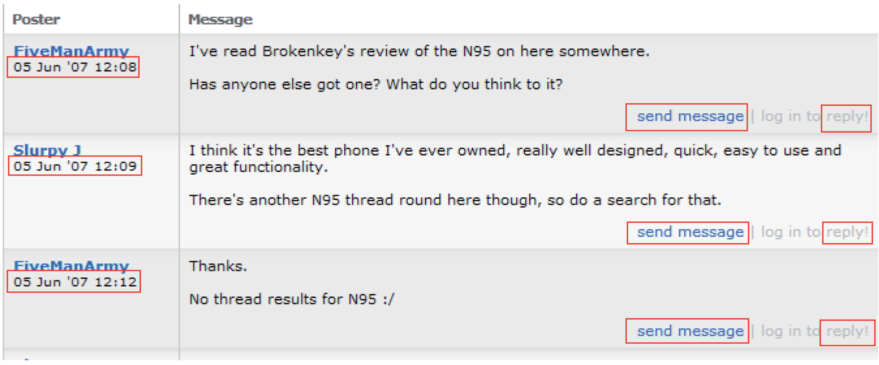
4.3. Domain Pattern - Discussion (Forum)

Problem: How to hold a discussion about a certain topic? How to show a summary of comments and opinions?

Context: Social field, community sites, blogs, etc. Discussions about product and services selling. Review discussions. News story discussion.

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

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 pattern is often together with the *Login* pattern. If the pattern is on a selling product website there are usually *Purchase possibility*, *Price information* patterns. The pattern can be alone on the page. In other case there is also the *Something to read* pattern. In different domains the pattern can be together with patterns *Review*, *News*, etc. See Fig. 5.



Poster	Message
FiveManArmy 05 Jun '07 12:08	I've read Brokenkey's review of the N95 on here somewhere. Has anyone else got one? What do you think to it? send message | log in to reply!
Slurpy J 05 Jun '07 12:09	I think it's the best phone I've ever owned, really well designed, quick, easy to use and great functionality. There's another N95 thread round here though, so do a search for that. send message | log in to reply!
FiveManArmy 05 Jun '07 12:12	Thanks. No thread results for N95 :/ send message | log in to reply!

Fig. 5. *Discussion* pattern.

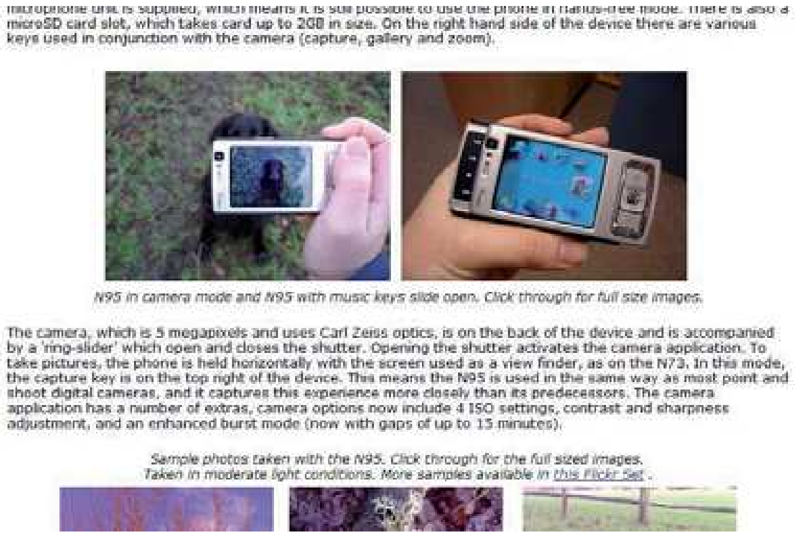
4.4. Structural Pattern - Something to Read

Problem: How to lucidly write text on the web page?

Context: The pattern is used often and regardless of the domain.

Forces: A fragment occupying almost a whole page. There are usually longer continuous paragraphs. If the text is long there can be a short heading among some paragraphs. Inside paragraphs there can be a few links or images.

Solution: It is simply implemented using line spacing and headings of a certain level. See Fig. 6.



microphone unit is supplied, which means it is still possible to use the phone in hands-free mode. There is also a microSD card slot, which takes card up to 2GB in size. On the right hand side of the device there are various keys used in conjunction with the camera (capture, gallery and zoom).

N95 in camera mode and N95 with music keys slide open. Click through for full size images.

The camera, which is 5 megapixels and uses Carl Zeiss optics, is on the back of the device and is accompanied by a 'ring-slider' which open and closes the shutter. Opening the shutter activates the camera application. To take pictures, the phone is held horizontally with the screen used as a view finder, as on the N73. In this mode, the capture key is on the top right of the device. This means the N95 is used in the same way as most point and shoot digital cameras, and it captures this experience more closely than its predecessors. The camera application has a number of extras, camera options now include 4 ISO settings, contrast and sharpness adjustment, and an enhanced burst mode (now with gaps of up to 15 minutes).

Sample photos taken with the N95. Click through for the full sized images.
Taken in moderate light conditions. More samples available in this Flickr Set .

Fig. 6. *Something to read* pattern.

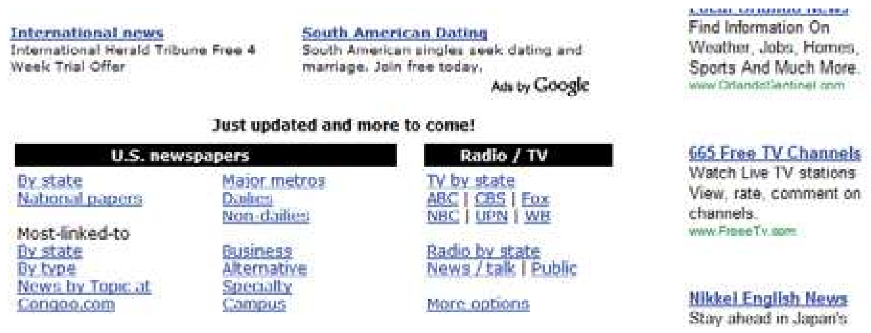
4.5. Structural Pattern - Link List

Problem: How to lucidly show list of links to related pages?

Context: The pattern is used often and regardless of the domain.

Forces: The page fragment with links. Each link is usually in the form of intelligible text within the scope of single sentence (few words). Each link can be appended with a short text or URL address. Each link is on a single row.

Solution: There is usually implementation with single continuing rows or a similar strategy leading to the same-looking result (for example using enumerations, lists, table layouts). See Fig. 7.



International news
International Herald Tribune Free 4 Week Trial Offer

South American Dating
South American singles seek dating and marriage. Join free today.

Ads by Google

Find Information On Weather, Jobs, Homes, Sports And Much More.
www.OrlandoSentinel.com

Just updated and more to come!

U.S. newspapers

By state
National papers

Major metros
Dailies
Non-dailies

Most-linked-to
By state
By type
News by Topic at Congoo.com

Business
Alternative
Specialty
Campus

Radio / TV

TV by state
ABC | CBS | Fox
NBC | UPN | WB

Radio by state
News / talk | Public

More options

665 Free TV Channels
Watch Live TV stations View, rate, comment on channels.
www.FreeeTv.com

Nikkei English News
Stay ahead in Japan's

Fig. 7. *Link list* pattern.

5. PAGE SIMILARITY

5.1. Human view

With the assumption that it is possible to detect web patterns automatically on the web pages it is possible to describe each page with the patterns. Using pattern names the description may look like this example: “The page contains the *Price information*, the *Purchase possibility* and the *Special offer*. There are also *Technical details* and the *Discussion* at the bottom.” Using such a description the user does not know which product is presented on the page but the user as well as the page designer can imagine how the page looks like. So the group of patterns characterizes a relatively wide set of pages which has similar intent (and belongs to the same domain). Such a group of patterns we call a page profile.

5.2. Technical View

Pattern detection. In the paper [9] we provided a general description of algorithm based on so called Gestalt principles (proximity, closure, similarity, continuity – see [14]). The algorithm can detect domain patterns on web pages with a high relevance (about 80%). As described above the algorithm makes the content segmentation and segment extraction of a web page, segment evaluation and evaluation of relevance (weight) of the found pattern. In page segmentation we work with dynamic generated dictionaries of patterns containing frequently used words and data types in the pattern context. On Fig. 8 there is a simplified representation of a pattern in XML used for setting up the algorithm. The whole system for pattern detection is opened even for example to various language versions.

For the detection of structural patterns we use specialized algorithms. In the HTML code extraction phase we preserve only few elements for example links and paragraphs. These elements are foundations for algorithms design.

```
<PATTERN>
  <ID>100</ID>
  <NAME>Price information</NAME>
  <PROXIMITY>8</PROXIMITY>
  <BASE_WEIGHT>1</BASE_WEIGHT>
  <PROMINENCE_WEIGHT>1</PROMINENCE_WEIGHT>
  <CLOSURE_WEIGHT>2</CLOSURE_WEIGHT>
  <SIMILARITY_WEIGHT>0.25</SIMILARITY_WEIGHT>
  <CONTINUITY_WEIGHT>0</CONTINUITY_WEIGHT>
  <MAIN_KEYWORDS>
    <WORD>$</WORD>
    <WORD>price</WORD>
    <WORD>vat</WORD>
    <WORD>prices</WORD>
  </MAIN_KEYWORDS>
  <COMPLEMNTARY_KEYWORDS>
    <WORD>availability</WORD>
    <WORD>shopping</WORD>
    <WORD>cart</WORD>
    <WORD>item</WORD>
  </COMPLEMNTARY_KEYWORDS>
</PATTERN>
```

Fig. 8. Representation of Price information pattern.

Page representation. In vector model, a document (Web page) W_j is represented as a vector w_j of pattern weights, which record the extent of importance of the pattern for the Web page. To portrait the vector model, we usually use an n × m pattern-by-page matrix A, having n rows – pattern vectors $p_1 \dots p_n$ – where n is the total number of terms in collection **C** and m columns – page vectors $w_1, \dots, w_m$, where m is the size of collection **W**.

The dimension is never too big. The sources for patterns are catalogs **C**. These catalogs are not too large. However, the web patterns are not described with regard to automatic detection. This is a reason why we have our own catalog and describe web patterns according to the patterns described in our paper.

Weights of patterns are obtained as the result of the pattern detection algorithm application. This value means level of certainty whether a searched pattern was detected on the Web page (a pattern weight). There are several ways how to search for relevant documents. Generally, we can compute some L_n metrics to represent the similarity of pattern vectors. However, in text retrieval better results can be obtained by computing similarity, usually using the cosine measure:

$$\mathrm{Sim}(p,q) = \frac{p \cdot q}{\| p \| \cdot \| q \|}$$

6. EXPERIMENT: REORDERING

With the chosen set of users we wanted to prove our approach. We were interested in whether the automatic web pattern detection on web pages may be useful for the users. It may help then with orientation while searching through a vast amount of web pages. We designed an experimental web interface which is similar to a common search engine (www.pattrio.net). In the user interface the user could write a query in a common way. For the purpose of this experiment we use data provided by the Czech search engine *www.jyxo.cz*. After that the search engine returned a set of 100 pages including page text content. The set was analyzed. For each web page the vector was computed which represents the page and which aggregated the best segments from searched web patterns. Then the page set was displayed to the users in the original order from the search engine. The vector (the page profile) was displayed as a tie-on label (see Figure 1). Altogether we used 13 domain and 5 structural patterns. By clicking on the tie-on label the result set would be reordered according to the profile of the selected page and less according to the original order. The selected page was on the first position in the result set and the least similar page was moved to the end of the set. In this reordering there were many pages from the second half of the original order moved to the first ten of pages.

The users (students and teachers) were instructed that by clicking on the tie-on label of the web page the result set will be reordered. The interaction between the user and the result set was completely recorded for further analysis. The users were asked to answer three questions. The answer could be only *Yes* or *No*. There were 34 users answering.

1. Is the information about the searched page which was being displayed useful for selecting the page? There were richer snippets below the title and the tie-on label on the right side.
2. Does the displayed information correspond with the page content?
3. Is the reordering, according to the tie-on label, helpful in searching?

We got 21 *Yes* answers on the first question. For the second question there were 34 users answering *Yes* with one note. The note is that what was written on the tie-on label was mostly O.K. but some web patterns presented on the page were sometimes not detected. For the last questions we got 26 positive answers with a note that the reordering was useful but only for pages where web patterns were detected. The most common comments were:

1. 1. It is good that the pages which I am not interested in at the moment are moved to the end of the result set.
2. There are too many pages with no web pattern detected and the reordering according to their profile does not bring the expected result (in the meaning of similarity with selected page).
3. The pages are not always ordered as I expected.
4. Sometimes the result set is reordered toward product selling but I get pages about a different product than I wanted.

The third criticism relates with use of cosine measure. The best results can not always be reached with this measure. We prepared experiments with different methods for computing similarity.

According to the last point we added links with words from a title below each title. By clicking on the link the word was added to the query (see Figure 1). The idea that the title usually contains words important for the page content can be considered as a web pattern.

7. EXPERIMENT: PROFILES AND CLUSTERS

In the previous experiment we worked with a profile of each page in the reordering of the result set. The goal of the second experiment was to find whether some static or common profiles exist. The profiles can be used for reordering of the resulting set. The intuition is showing that there should be profiles for internet shops, advertising and forums for sharing experience. For this experiment we used the recorded pages from querying products. The analysis was performed on 23,422 pages. Each page contained at least one domain pattern from the selling products domain (altogether it was 9 domain patterns). We wanted to visualize the analysis results so we used the SOM method. The self-organizing map (SOM) is a type of neural network. This network is trained using unsupervised learning and produces low dimensional representation of the training samples while preserving the topological properties of the input space. The model was first described by Teuvo Kohonen and it is also known as a Kohonen map [8]. On Fig. 9 there is the Kohonen map displayed with numbered clusters.

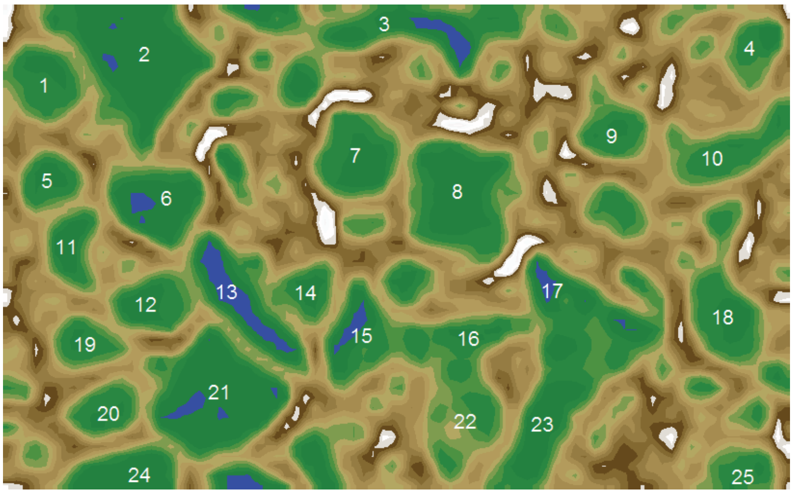


Fig. 9. SOM – web pages from selling product domain.

In each cluster there are pages which:

1. Contain mentioned web patterns in the table (moreover with value greater than 0.6).
2. Do not contain other web patterns (with weight lesser than 0.3)

The clusters are numbered and described in Table I. The web patterns names are abbreviated. There are four profile types represented by searched cluster:

1. Selling: 1, 2, 4, 6, 7, 8, 9, 10, 18, 20, 21, 24, 25 (6570 pages)
2. Information: 3, 16, 17, 22, 23 (3,570 pages)
3. Advertising: 5, 11, 12, 19 (1,050 pages)
4. Another clusters: 13, 14, 15 (2,310 pages)

In all mentioned clusters there are approximately 13,500 pages. For approximately 11,200 pages it is possible to assign their type (*Selling, Information, and Advertising*).

On approximately 2,300 pages it is not possible to say whether they are from a different domain or it is an error in the web pattern detection. Probably the pages are from a different domain than the selling product domain. They may be pages for example, from a tourism domain which requires the use of different web patterns (patterns *Price* and *Login* can occur). Beside the discussed pages, there are approximately 10,000 pages which are not in any cluster. This is the case when there was detected a different combination of detected web patterns with different weights.

Table I. Found clusters.

#	Patterns	Pages
1	Price, Purchase, Special, Details	260
2	Price, Purchase, Special	1280
3	Review	1380
4	Price, Purchase, Special, Repayment, Details, Login	170
5	Price, Purchase, Special, Advert	200
6	Price, Special	630
7	Price, Purchase, Login	580
8	Price, Purchase, Special, Login	650
9	Price, Purchase, Details, Login	200
10	Price, Purchase, Special, Details, Login	340
11	Price, Special, Advert	270
12	Price, Advert	360
13	Price	1410
14	Price, Login	240
15	Login	660
16	Discussion, Login	340
17	Discussion	750
18	Price, Purchase, Special, Repayment	350
19	Price, Purchase, Advert	220
20	Price, Purchase, Details	260
21	Price, Purchase	1360
22	Discussion, Review, Login	300
23	Discussion, Review	800
24	Price, Purchase, Special	350
25	Price, Purchase, Special, Repayment, Details	140

8. CONCLUSION

We were focusing on the selection and the descriptions of the web patterns regard to the analysis of the web pages. The key factor for us is the human factor. We are convinced that using web patterns which are developed within interaction between users and web page designers is very useful. Our experiments show that there are two perspectives. The first perspective is purely technical. It is possible to extend meta-information about a web page with a page profile (coming from web patterns). The information should, be for example, used for search engines. The second perspective is user based. In our testing of web interface we tried to involve users in the process of searching and reordering the result set. Our experiments imply that the users understand the interface. Currently we prepare the next experiments using web pattern detection in web searching.

REFERENCES

[1] ALEXANDER, CH. 1977, *A Pattern Language: Towns, Buildings, Construction*. Oxford University Press, New York, USA.

[2] CHAKRABARTI S. 2003, *Mining the Web: Discovering Knowledge from Hypertext Data*. Morgan Kaufman Publishers.

[3] CHANG. CH.H., KAYED, M., GIRGIS, M.R., SHAALAN, K.F. 2006. A Survey of Web Information Extraction Systems. *IEEE Transactions on Knowledge and Data Engineering 18(10)*, 1411-1428.

[4] DEARDEN, A., FINLAY J. 2006. Pattern Languages in HCI: A critical review. *Human-Computer Interactio*n *21(1)*, 49-102.

[5] VAN DUYNE, D.K., LANDAY, J.A., HONG, J.I. 2002. *The Design of Sites: Patterns, Principles, and Processes for Crafting a Customer-Centered Web Experience*. Addison-Wesley Professional.

[6] IVORY, M.Y., MEGRAW R. 2005. Evolution of Web Site Design Patterns. *ACM Transactions on Information Systems 23(4)*, 463–497.

[7] KIYAVITSKAYA, N., ZENI, N., CORDY, J.R., MICH, L., MYLOPOULOS, J. 2006, Text Mining Through Semi Automatic Semantic Annotation. 143-154. In *Proceedings of 6th International Conference on Practical Aspects of Knowledge Management*. Vienna, Austria, November 30.

[8] KOHONEN T. 2000. *Self-Organizing Maps*, Springer.

[9] KUDELKA, M., SNASEL, V., LEHECKA, O., EL-QAWASMEH, E. 2006. Semantic Analysis of Web Pages Using Web Patterns. In *Proceedings of International Conference on Web Intelligence.* Hong Kong, 329-333.

[10] NIE, Z., WEN, J-R., MA, W-Y. 2007. Object-level Vertical Search. In *Proceedings of Conference on Innovative Data Systems Research.* Asilomar, CA, USA, 2007, 235-246.

[11] NIE, Z., MA, Y., SHI, S., WEN, J-R., MA, W-Y. 2007. Web Object Retrieval. In *Proceedings of International World Wide Web Conference*. Banff, Alberta, Canada, 81 - 90.

[12] PIVK, A. 2005. *Automatic Ontology Generation from Web Tabular Structures*. PhD thesis, University of Maribor.

[13] REIS, D.C., GOLGHER, P.B., SILVA, A.S., LAENDER, A.F. 2004. Automatic web news extraction using tree edit distance. In *Proceedings of International World Wide Web Conference*, New York, USA, 502-511.

[14] TIDWELL, J. 2006. *Designing Interfaces: Patterns for Effective Interaction Design*. O'Reilly Media, Inc.

[15] YU, S., CAI, D., WEN, J-R., MA, W-Y. 2003. *Improving Pseudo-Relevance Feedback in Web Information retrieval Using Web Page Segmentation*, In *Proceedings of International World Wide Web Conference*. Budapest, Hungary, 203-211.

[16] *Beck Form*. http://c2.com/cgi/wiki?BeckForm, (October 13th 2007).

Václav Snášel
Department of Computer Science
Faculty of Electrical Engineering and Computer Science
VŠB -- Technical University of Ostrava
17. listopadu 15
708 33 Ostrava – Poruba
Czech Republic

Miloš Kudělka
Inflex, s.r.o.
Polívkova 10
779 00 Olomouc
Czech Republic

Ondřej Lehečka
Department of Computer Science
Faculty of Electrical Engineering and Computer Science
VŠB -- Technical University of Ostrava
17. listopadu 15
708 33 Ostrava – Poruba
Czech Republic

Received October 2007