User Profiling on a Pilot Digital Library with the Final Result of a New Adaptive Knowledge Management Solution

Karl Petrič*, Teodor Petrič**,
Marjan Krisper***, and Vladislav Rajkovič****

* Special library, Ministry of the Interior,
Štefanova ulica 2, 1000 Ljubljana, Slovenia, <karl.petric@gov.si>
** Department of German Studies, Faculty of Arts, University of Maribor,
Koroška cesta 160, 2000 Maribor, Slovenia, <teodor.petric@uni-mb.si>
*** Faculty of computer and information Sciences, University of Ljubljana,
Tržaška ulica 25, Ljubljana, Slovenia, <marjan.krisper@fri.uni-lj.si>
**** Faculty of Organisation Sciences, Department of informatics, University of Maribor,
Kidričeva 55a, 4000 Kranj, Slovenia <vladislav.rajkovic@fov.uni-mb.si>

Karl Petrič is an information librarian-scientist at the Special Library of the Ministry of the Interior, Slovenia. He received his Master's Degree in 2005, and his Ph.D. in 2008, both from the University of Ljubljana, Faculty of Computer and Information Science, Slovenia. His research interests include Knowledge Management, Data Mining, Text Mining, Knowledge Discovery, and Educational Information Systems. He is co-operating in projects on E-Archiving, Knowledge Intranet Portal, and in the preparation of university teaching materials for German Phonology.

Teodor Petrič is an Associate Professor of German Linguistics at the University of Maribor, Faculty of Arts. He received his Master's in German Linguistics in 1990, and his Ph.D. in Linguistics in 1995, both from the University of Ljubljana, Faculty of Slovenia. His research interests include Naturalness Theory, First and Second Language Acquisition Theory, Translation Tools, Study of Child Language, Text Mining, and Educational Information Systems. He is co-operating in projects on Language acquisition, Natural Linguistics, Phraseology, and Translation Documentation Systems.

Marjan Krisper is an Associate Professor of Information Systems at University of Ljubljana, Faculty of Computer and Information Science. He received his Master's in Information Systems Engineering from University of Ljubljana, Slovenia, in 1977, and his Ph.D. in Expert Systems from University of Belgrade, Yugoslavia, in 1989. His research interests include Electronic Business, Information Systems Development Methodologies, Mobile Applications, Information Systems Strategic Planning, Data Mining, and Expert Systems. He has been a project leader of various information systems development and other projects.

Vladislav Rajkovič is a professor of Management Information Systems, Faculty of Organizational Sciences, University of Maribor and Research fellow, Department of Intelligent Systems, Jozef Stefan Institute. He received a B.Sc. from the Faculty of Electrical Engineering, University of Ljubljana in 1970 and a M.Sc. from the Faculty of Electrical Engineering, University of Ljubljana in 1975. In 1987, he received a Ph.D. from the Faculty of Electrical Engineering and Computer Science, University of Ljubljana. Main Research Interests are information systems and their application for decision support, artificial intelligence methods for decision support, knowledge management for decision support etc.

Petrič, Karl, Petrič, Teodor, Krisper, Marjan, and Rajkovič, Vladislav. **User Profiling on a Pilot Digital Library with the Final Result of a New Adaptive Knowledge Management Solution.** *Knowledge Organization, 38(2)*, 96-113. 13 references.

ABSTRACT: In this article, several procedures (e.g., measurements, information retrieval analyses, power law, association rules, hierarchical clustering) are introduced which were made on a pilot digital library. Information retrievals of web users from 01/01/2003 to 01/01/2006 on the internal search engine of the pilot digital library have been analyzed. With the power law method of data processing, a constant information retrieval pattern has been established, stable over a longer period of time. After this, the data have been analyzed. On the basis of the accomplished measurements and analyses, a series of mental models of web users for global (educational) purposes have been developed (e.g., the metamodel of thought hierarchy of web users, the segmentation model of web users), and the users were profiled in four different groups (adventurers, observers, applicable, and know-alls). The article concludes with the construction of a new knowledge management solution called multidimensional rank thesaurus.

Received 3 May 2009; Revised 27 May 2009; Accepted 27 May 2009

1.0 Introduction

Digital libraries are not only systematically arranged collections of knowledge available via a computer communications medium, but are, in the final instance, information systems for managing knowledge able to suggest various important decisions (e.g., business decisions related to the educational process) and should additionally serve as state-of-the-art research platform for studying navigational patterns of users. Studying navigational behaviours (e.g., information retrieval, visits) of web users of digital libraries is one of the preconditions for devising new knowledge management solutions, which are adapted to various needs of users (internal users, external users, etc.). There is an insufficient number of studies related to navigational behaviours of web users and determining user profiles in web (digital) libraries, which is a great deficiency for quality of new knowledge management solutions, as the users are a sort of a primary key for operation of any user-oriented application system.

1.1 Approach and framework

We propose a user-oriented approach in solving the issues of studying users of digital libraries (Mayr 2004). The essence of the approach is the framework for studying navigational (search) patterns of users discussing socio-technical aspects with the purpose of devising a model of a new knowledge management solution called multidimensional rank thesaurus. The framework for studying navigational patterns and, in the next stage user profiles, consists of the following elements (figure 1 and description).

Figure 1 shows the framework for studying navigational patterns of web users with the final result of a new adaptive knowledge management solution called multidimensional rank thesaurus. In this work we will point out the following four steps (a – d):

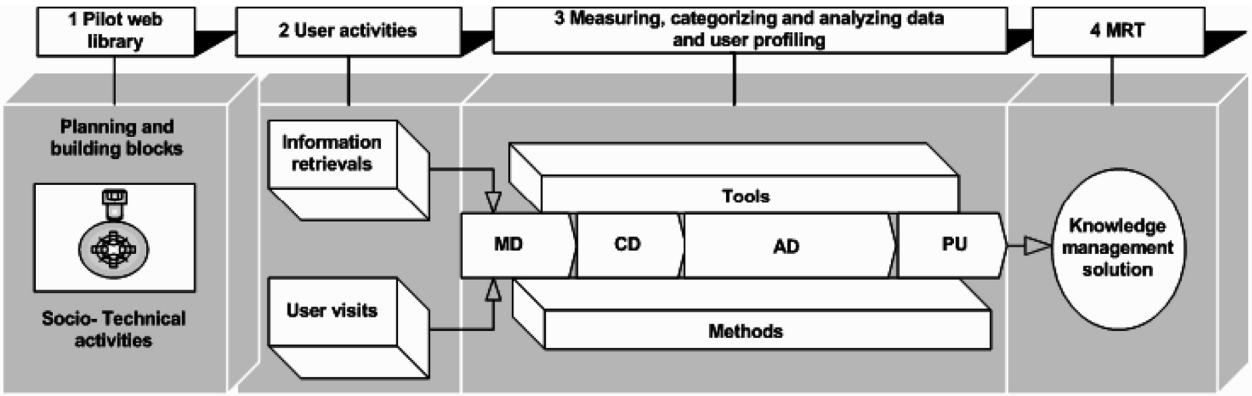

Figure 1. Framework for studying navigational patterns of users with the final result of a new adaptive knowledge management solution

a. First step, alias "pilot web library"

Devising of a pilot version of a web library on the basis of information needs of users in a secondary high school library (socio-technical approach):
- monitoring user activities in a classic secondary school library,
- monitoring library users – determining the rhythm of their activities and identifying regular users,
- monitoring borrowing, recording information needs of users, etc. (study period six months).

After the pilot web library was built, we measured the number of web user visits and their information retrievals for another six months. On the basis of this useful information feedback, we added links on the subpages, and, in the year 2003, we started the real measurements. The pilot web library was built with free and open source web services and tools. It was based on the studied classic library with free access to materials. Both the materials and the pilot web library, including all of its subpages, were systematically arranged under the Universal Decimal Classification system (hereinafter the UDC). Software tools (e.g., internal search engine, trackers of user activity, discussion forum, web survey) by which user activities could be monitored directly or indirectly were included.

b. Second step, alias "web user activities"

Data capture and classification: in the period between January 1st, 2003 and January 1st, 2006, web visits and information retrievals made by web visitors in the in-house devised web library were collected (Pabarskaite and Raudys 2007). The obtained data were promptly and intellectually classified by using the UDC (only the main classifier of each class was taken into account). We developed many classification rules on information retrievals, and, in some cases, we also took into account the visiting frequencies (in the case information retrievals could not be exactly classified, reconstructions of user visits on the pages of the UDC web library were made). The obtained data were taken from the free Pico search engine panel (file format was .html) and exported into Excel every day. In three years, we collected 13,613 information retrievals. The mainstream of the targeted audience were elementary and high school students. In addition to search criteria, we also counted the number of visits per web library subpage (already UDC classified) with the free Nedstat tracker tool (nowadays Motigo). These data can later be used for additional detailed analysis of information retrieval and visits (Kaushik 2007).

c. Third step, alias "process of measuring (MD), categorizing (CD), analyzing data (AD) and user profiling (PU)"

Data planning and preparation was based on the CRISP methodology which includes the following elements (Chapman et al. 2000): familiarity with the field of research, understanding data, preparation of data, data modeling, data evaluation and data construction (Berry and Linoff 2004). The planned and prepared data enabled us for in depth data analysis (AD), the basis for user profiling (PU).

d. Fourth step, alias "the new adaptive knowledge management solution or multidimensional rank thesaurus (MRT)"

The term multidimensional rank thesaurus (MRT) is relatively new in this context. On the internet and other data collections (Web of Science, INSPEC etc.), this term does not appear very often except for only a few hits in connection with transport. MRT in our paper is defined as a retrieval system, where the data are organized in a thesaurus consisting of two or more integrated thesauri with different dimensions. In our case, there are three dimensions (the institutional program, ranking keywords, and user profiles).

1.2 Methodology and methodological tools

The research method includes total observation of randomly chosen web users in the above-mentioned three-year period. The visits and the information retrieval activities of the users in the pilot web library were observed on a daily basis, followed by UDC classification of the data. As mentioned, the data classification was conducted intellectually. In our UDC classification of the data, only the main UDC numbers were taken into account. In difficult cases, the navigation of the web users from one to another subpage was additionally used to classify information retrievals: e.g., a web visitor conducted an information retrieval with the keywords "social psychology," but instead of visiting the subpage on social sciences

(with the UDC number 3), he or she visited the philosophy and psychology subpages (with the UDC number 1). Based on the navigation dynamics, this information retrieval (and similar ones) was put into the UDC category number 1 afterwards, handled according to our data analysis procedures. For data analysis, the following software tools were used: MS Excel 2003, AntConc 3.1.302, Orange Canvas 0.9.6.2, CBA association Rules 1.0.

1.3 Research hypotheses and research questions

Based on the accomplished information retrieval categorization, the following research questions and research hypotheses were deducted.

1.3.1 Research hypotheses

1.) In the long run, the information retrieval pattern of the visitors of the existing pilot web library does not change and, therefore, displays independency of time.
2.) According to the alternative hypothesis, the information retrieval pattern of the visitors of the existing pilot web library does change during a longer period of time.

1.3.2 Research questions

1.) Which dominating information retrieval pattern(s) can be identified? Does the information retrieval pattern of the pilot web library users change over time?
2.) Which information retrieval patterns of the pilot web library users reappear constantly?
3.) Is it possible to create profiles of the pilot web library users according to the engaged methods and tools?
4.) How many different user profiles are to be distinguished?
5.) Is it possible to build an application based on our knowledge of the user profiles and capable of adapting to their information needs?

In this article, we will focus on step 3 (AD – Data analysis, PU – User profiling) and step 4 (knowledge management solution – multidimensional rank thesaurus MRT). Previous steps (concerning the pilot web library, user activities and two parts of step 3 – MD and CD) have been described briefly in subsection 1.1 (Approach and framework).

2.0 Data analysis (AD)

Here we describe how we arrived at different user profiles based on our data analysis. At first, we will determine the power law from information retrievals, and, in the next stage, we will discover and extract new knowledge.

2.1 Determining the power law

We calculated the logarithms of the frequencies of information retrievals, which were classified in different UDC areas. After this procedure, we compared the data from each year with all years in the whole period (2003, 2004, and 2005).

Table 1 and Figure 2 present three tests and the real power law of data. The three tests for establishing the power law were positive (log N2003, log N2004, and log N2005), so we could start the real power law of 13,616 information retrievals between the years 2003 and 2006 (log Ntogether), which were realized on the internal search engine of the in-house constructed pilot web library. The frequencies were calculated as C = log N (C stands for constant and N for frequency). At the x- axis, we have ranks of UDC areas, and, at the y- axis, we have values for log N (e.g., see rank 1 UDC 9 with the value 3.4951, rank 2 for UDC 8 with the value 3.2826, etc.). The method for determining the power law on the basis of collected web data on information retrievals of users in the pilot web library and their classification enables proving the universal pattern of user behaviour regarding their areas of interest, which remains constant even in a longer term (Bak 1997). In this research, we got a relative universal pattern of information retrieval bevahiour of web users, which is constant even in a longer term. So the first hypothesis (see subsection 1.3.1) was confirmed, which enabled us to answer the first research question.

In order to answer the third research question (see subsection 1.3.2), we have to point out that almost all user activities (information retrievals and visits) were guided by high school educational programmes (e.g., information retrieval for preparing a seminar or investigation works on topics like life and work of composers, writers, countries, osmosis, mechanics, religious communities, informatics, stress in schools, antic philosophy). So it is no surprise that the mentioned information retrievals mostly contributed to the universal pattern of information retrieval behaviour, which, of course, did not change even over a longer period of time. The high school programmes also determined

Rank (R)	log N together	UDC	Rank (R)	log N 2003	Rank (R)	log N 2004	Rank (R)	log N 2005
1	3,4951	9	1	2,8971	1	3,1000	1	3,0354
2	3,2826	6	2	2,7419	3	2,8645	2	2,8122
3	3,2702	5	4	2,6590	2	2,9053	4	2,7917
4	3,2620	3	3	2,7300	4	2,8299	3	2,8000
5	3,0990	8	5	2,4942	5	2,7348	5	2,6201
6	3,0626	7	6	2,4330	6	2,7168	6	2,5717
7	2,9274	1	9	2,3385	7	2,5211	7	2,4941
8	2,9079	0	7	2,4116	9	2,4624	9	2,1553
9	2,8195	2	8	2,3856	8	2,4713	8	2,4330
X	X	4	X	X	X	X	X	X

Table 1. Ranking UDC areas and logarithms of frequencies for years 2003, 2004, 2005 and all together

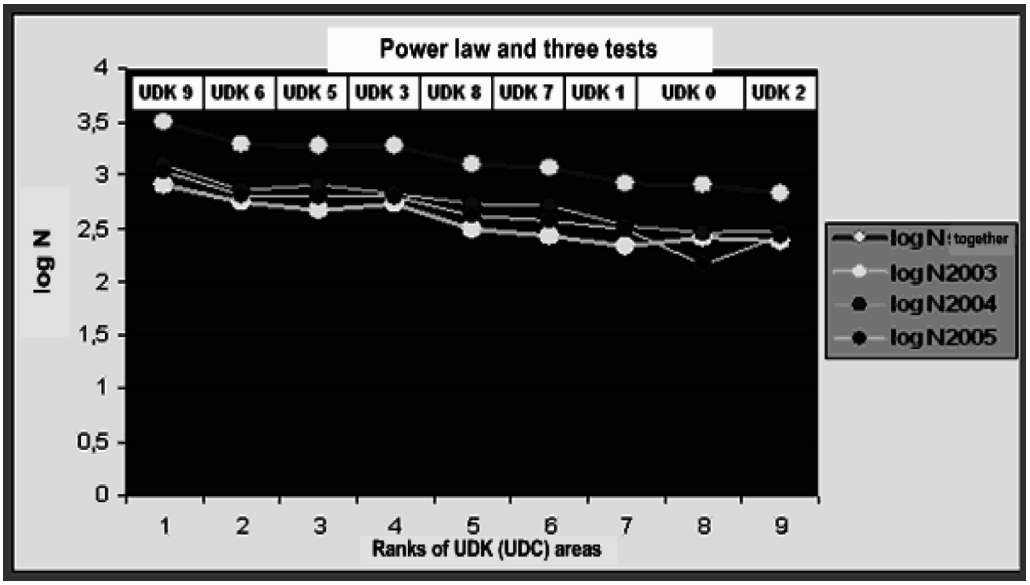

Figure 2. Power law and three tests

which keywords were likely to receive high ranks (e.g., biographies for seminar works), and, therefore, they had strong impact on the organization of the data in the MRT. Demonstrating such search behavioural patterns (see table 1 and figure 2) is the platform for developing the metamodel of thought hierarchy of web visitors and searchers as well as for subsequent discovering of laws, and in the next stage discovering of new knowledge. For better understanding, we will cut the big picture into two pieces.

Figure 3 shows the metamodel part 1 of thought hierarchy of web users with the distinction of their educational interests, which was accomplished from the results of the web user analysis with the power law method. Figure 3 presents web users' respective internal search engine users main tendencies, which are as follows:

a. Tendency to use a tool for quick finding of information and problem solving. Problems are divided in different groups as applicative, information, navigation, orientation, learning, ethical, and hybrid.
b. Every search engine user has his mission, vision, and aim. He must follow several rules, which are navigational and also of content nature. Some

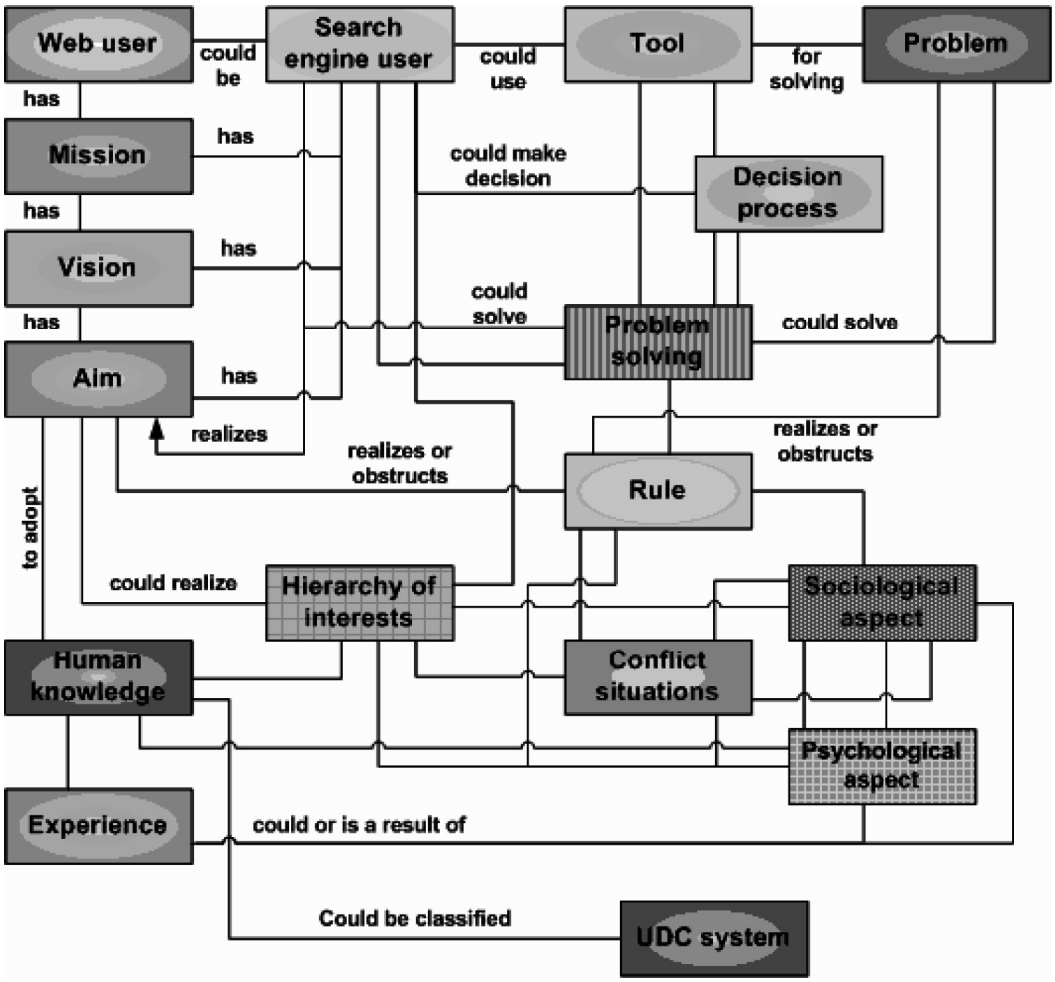

Figure 3. Metamodel part 1 of thought hierarchy of web users

rules could stimulate the realization of aims, but, on the other hand, there are some rules, which could block the user reaching his or her aims. Some rules are suitable for solving real world problems, but some specific rules obstruct the process of problem solving. Rules are also extremely involved in classification systems; in our example, it is the UDC.

c. The main aim of our (internal) search engine users is to solve a problem, to satisfy some information needs and finally to get some new (applicative) knowledge (see on figure 3 presented as human knowledge).

d. Data in the pilot UDC digital library are organized in UDC areas classifying all known human knowledge with ordinal numbers from 0, 1, 2, 3, 5, 6, 7, 8 to 9 (UDC 4 has the function of measuring information retrieval noise and in the conventional UDC system the UDC 4 group is empty without usage). Let us now turn to part two of the metamodel.

In figure 4, the thought hierarchy of the (internal) search engine users with respect to their educational interests is presented. The hierarchy was deducted from the power law (before data cleaning) and the analysis of information retrievals (after data cleaning). Figure 4 will not be described in detail, but we would like to point out that it shows the rank changes of the UDC areas before and after data cleaning. Data cleaning on the basis of a specially prepared stop list of words, the computation of the keyness factor, and the extraction of the most relevant keywords in the information retrieval corpus were carried out with the software tool *AntConc.* After data cleaning, UDC 9 area, which described the topics of world and human development, remained on first rank. Due to data cleaning two areas, UDC 7, which described the human knowledge about sensitive expressions, culture, and aesthetics, and UDC 6, which described the human knowledge of health and applicable things, exchanged their rank positions. UDC 6 area was demoted from the second rank be-

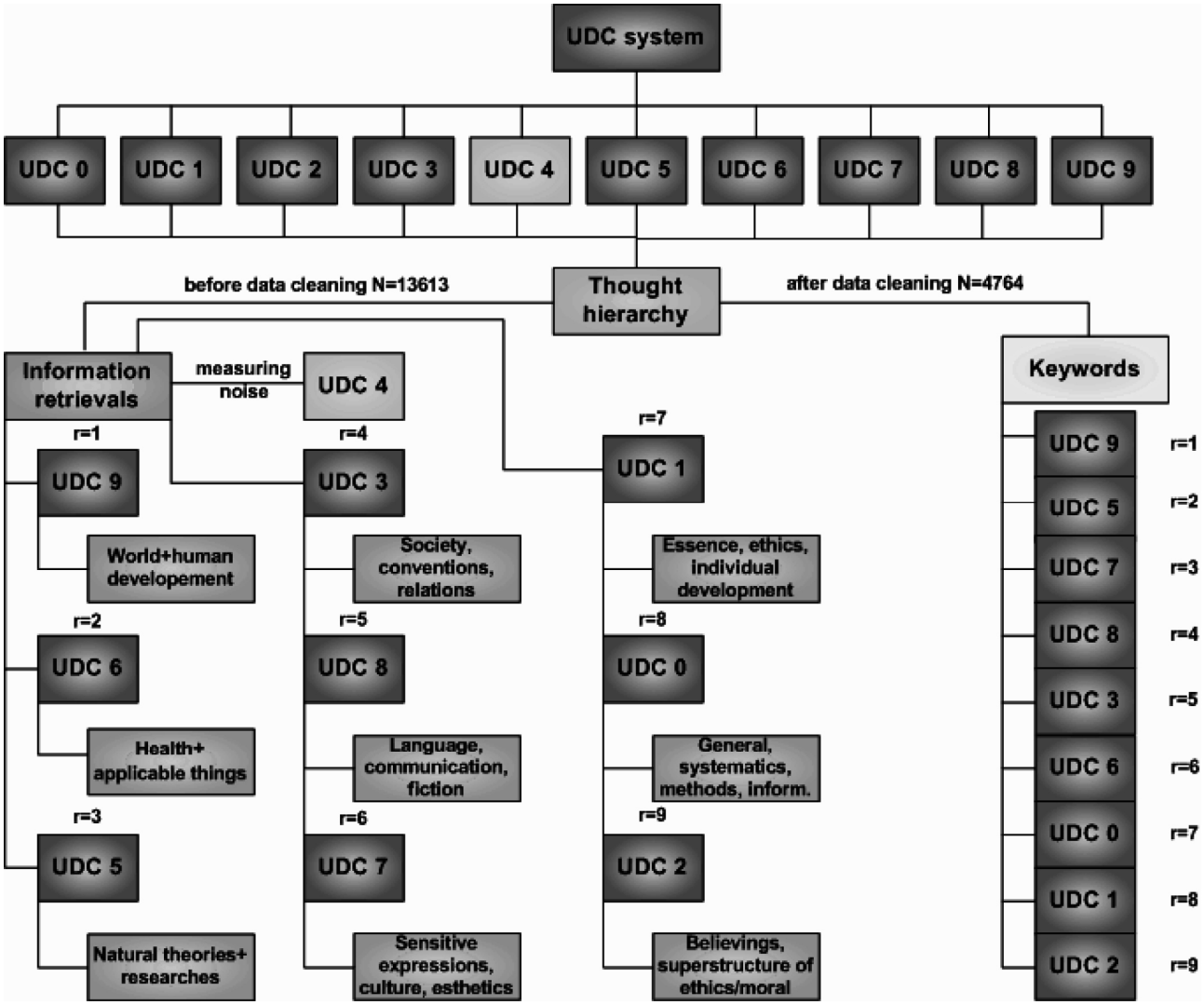

Figure 4. Metamodel part 2 of thought hierarchy of web users

fore data cleaning (reached by the number of information retrievals) to the sixth rank after data cleaning (due to the impact of the keyness factor). On the other hand, UDC 7 area was promoted from the sixth rank before data cleaning to the third rank after data cleaning. UDC 5 area was promoted from the third to the second rank. The exchange of the rank positions of UDC 6 and 7 displays the striking contrast between the impact of the information retrieval frequency and the keyness factor for UDC 6 area: the keyword factor pointed out that a certain amount of keywords used in the information retrievals were less relevant. On the other hand, the computed keywords from information retrievals for UDC 7 area were much more relevant than for the UDC 6 area. This insight is fundamental for the later procedure of user profiling, a crucial component of restoring and later monitoring of a multidimensional rank thesaurus. Certain methods in discovering laws in data and text (according to Konchady 2006, the association rules method, the hierarchic clusters method, the keyness factor determination) can be used to detect specific patterns and, in the next stage, to obtain new knowledge (e.g., profiled users), which subsequently leads us to the construction of a new knowledge management solution called multidimensional rank thesaurus, adapted to the users' interests.

2.2 Visual programming and analyzing of data

Figure 5 shows the procedure for visual programming of input (see figure 5: File of information retrievals N=13613) and output data (see figure 5: containing several visualizing techniques e.g. Scatterplot, Sieve

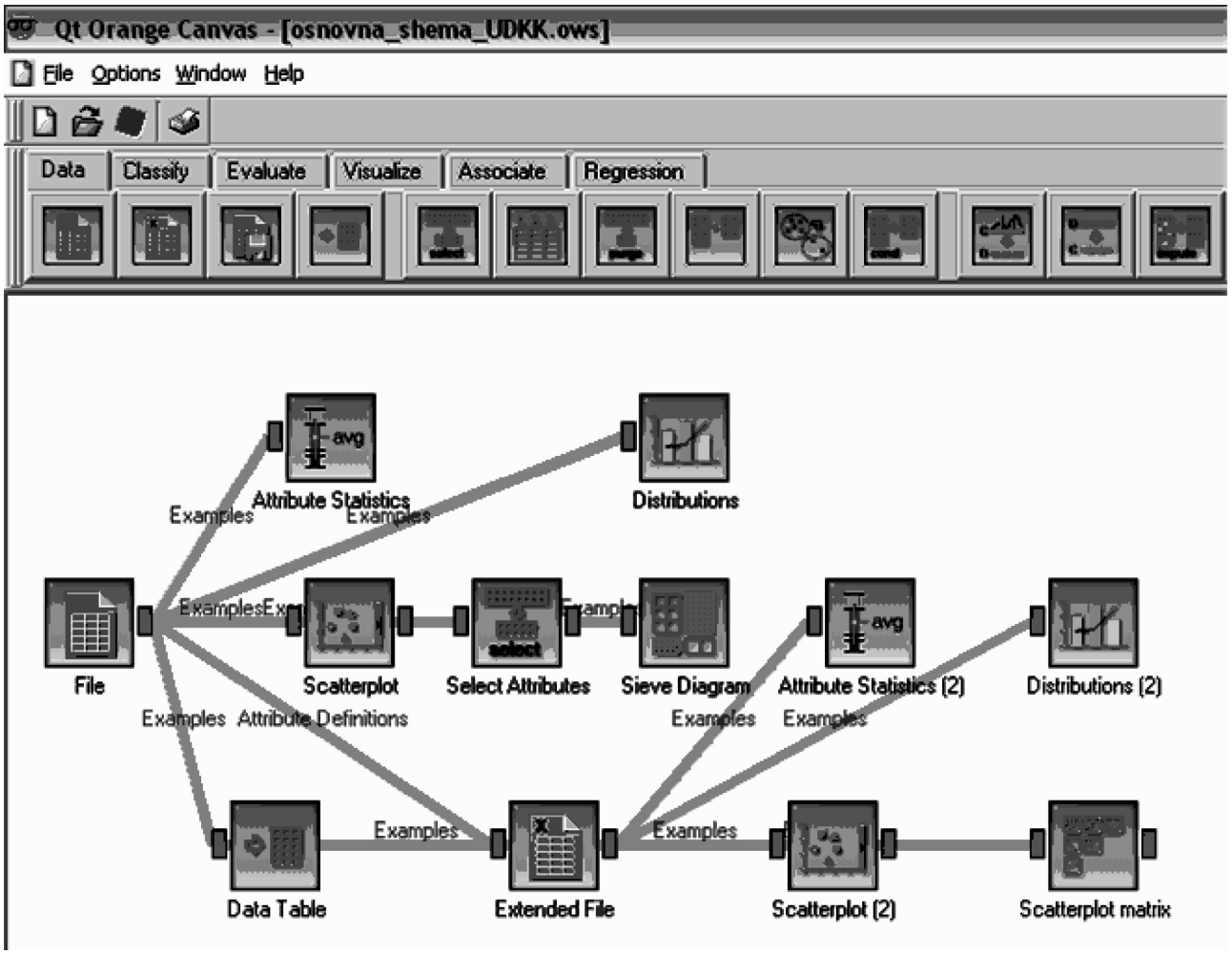

Figure 5. Visual programming of data (the basic scheme)

Diagram, Distributions). We compared the structured data between the classified information retrievals into UDC areas (X – axis) and the frequencies of them (Y – axis). Especially the Sieve Diagram shows us the occupied surfaces of different UDC areas (e.g., the biggest occupied surface was reached by UDC 9 area). The second input (in figure 5, it is the Extended File) we compared the ranks of UDC areas and the keyness factor of the relevant keywords inside the UDC areas (see Scatterplot 2). The scatterplot showed us the Pareto distribution of relevant keywords (N=4764), i.e., the Zipf distribution from another perspective. This means that a small number of relevant keywords contribute much more to the whole than the great majority of keywords. The most relevant keywords came from UDC 9 area (e.g., famous persons, history, countries). Two very important keywords came from UDC 1 area (the area of essence, ethics, and individual development/psychology: stress and schizophrenia). This part of the examination was very important for the later determination of association rules and hierarchical clusters of relevant keywords in the corresponding UDC areas by which we could create an educational user interest profile (we will describe this later).

Figure 6 shows us the visual programming procedure of data for discovering association rules and hierarchical clusters. This part of all procedures was one of the most crucial for establishing educational interests of user profiles. Only the global results will be pointed out in this paper. Association rules were obtained from two parallel procedures of numerical and of text data (Mccue 2007). The strongest association rule for numerical data eliminated all keywords with a keyness factor less than K and a frequency less than N (support = 0.960; confidence 0.984: $N \leq 19.80 --> K \leq 15.404$) and provides us with the most important keywords, which are on the interval of $K \geq 15.404$ and $N \geq 19.80$. This enabled us to carry out the association rules of the most important keywords, which were classified according to UDC ar-

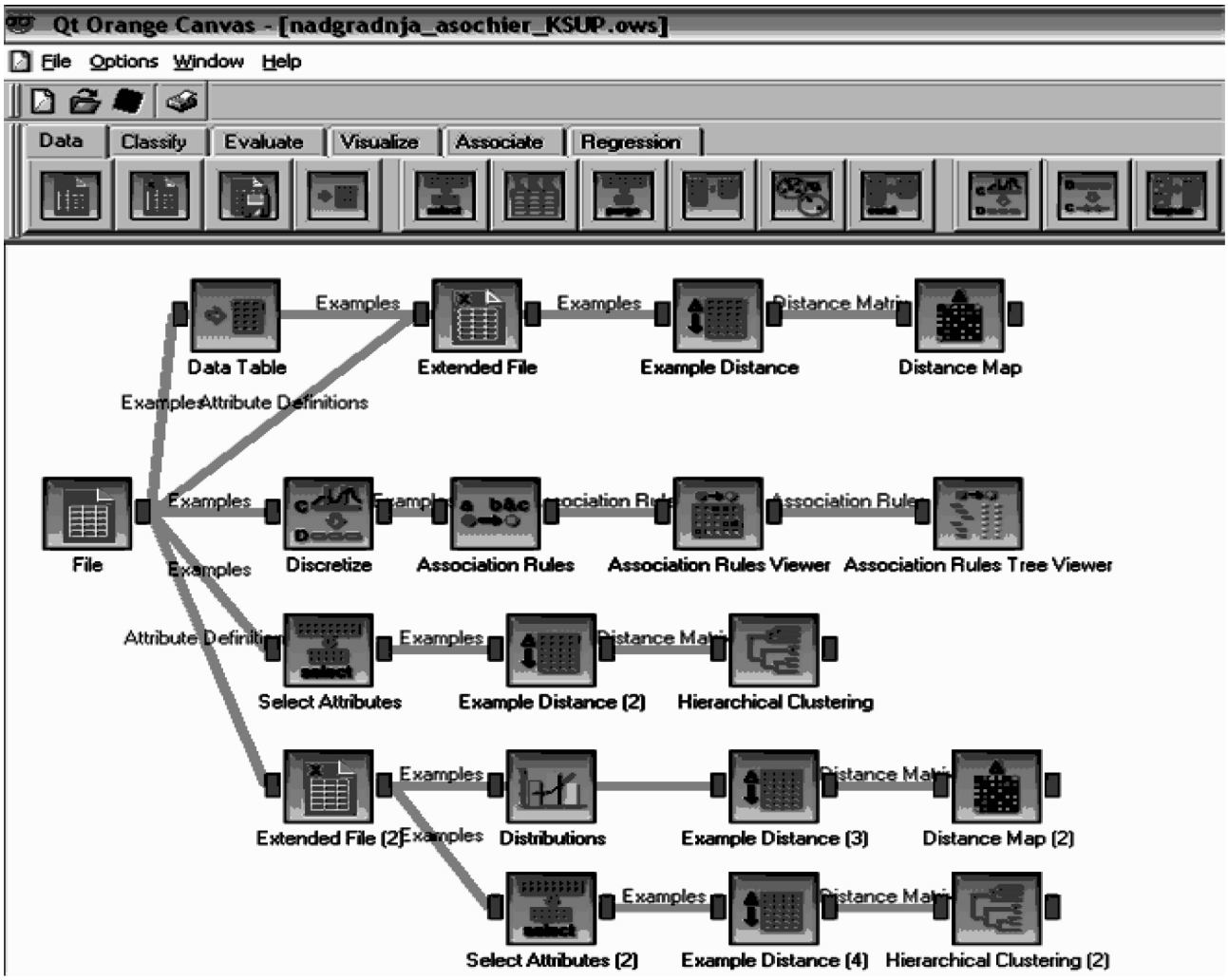

Figure 6. Visual programming of data for discovering association rules and hierarchical clusters

eas. The main layer for the association rules lay within the UDC areas 9, 8, and 7. In these three UDC areas, the closest connections between the relevant keywords could be found. This association rule pointed out one of the most prominent user groups (later on termed as "adventurers"). This group of users (and others) will be described after the procedure of hierarchical clustering and segmentation of web users (Mena 2003). For better understanding of this topic, the following picture is added.

Figure 7 shows the strongest association rule on the basis of keyness factor (K), rank (r), frequency, and UDC (UDK), described above. The left part of figure 6 compares the rank *r* (X-axis) with the keyness factor *K* (Y-axis) in order to find out the strongest keywords. The right part of figure 7 compares the keyness factor *K* (X-axis) with the frequency *N* of those keywords (Y-axis). Due to this relation, we obtained those keywords which are closer connected with the strongest keywords (see the left part of figure 6). The strongest keywords are visible in the upper part of figure 7 (e.g., history (zgodovina), stress (stres), Bach, UDC (UDK), Vivaldi, Haydn, Beethoven).

Table 2 shows the comparison between the percentages of the most relevant keyness factors (K%) and user visits (N%), which were classified into the given UDC classes 0, 1, 2, 3, 5, 6, 7, 8, and 9. The UDC values for user visits and keyness factors were converted into percents.

In the next step, clusters were computed with the Euclidian distance algorithm implemented in the program *Orange Canvas*. We identified four different clusters, which were as follows:

Cluster 1: UDC 9, 8, and 7 (common denominators: leisure, recognizing new things, inquiring about new topics, places and peoples, etc.)

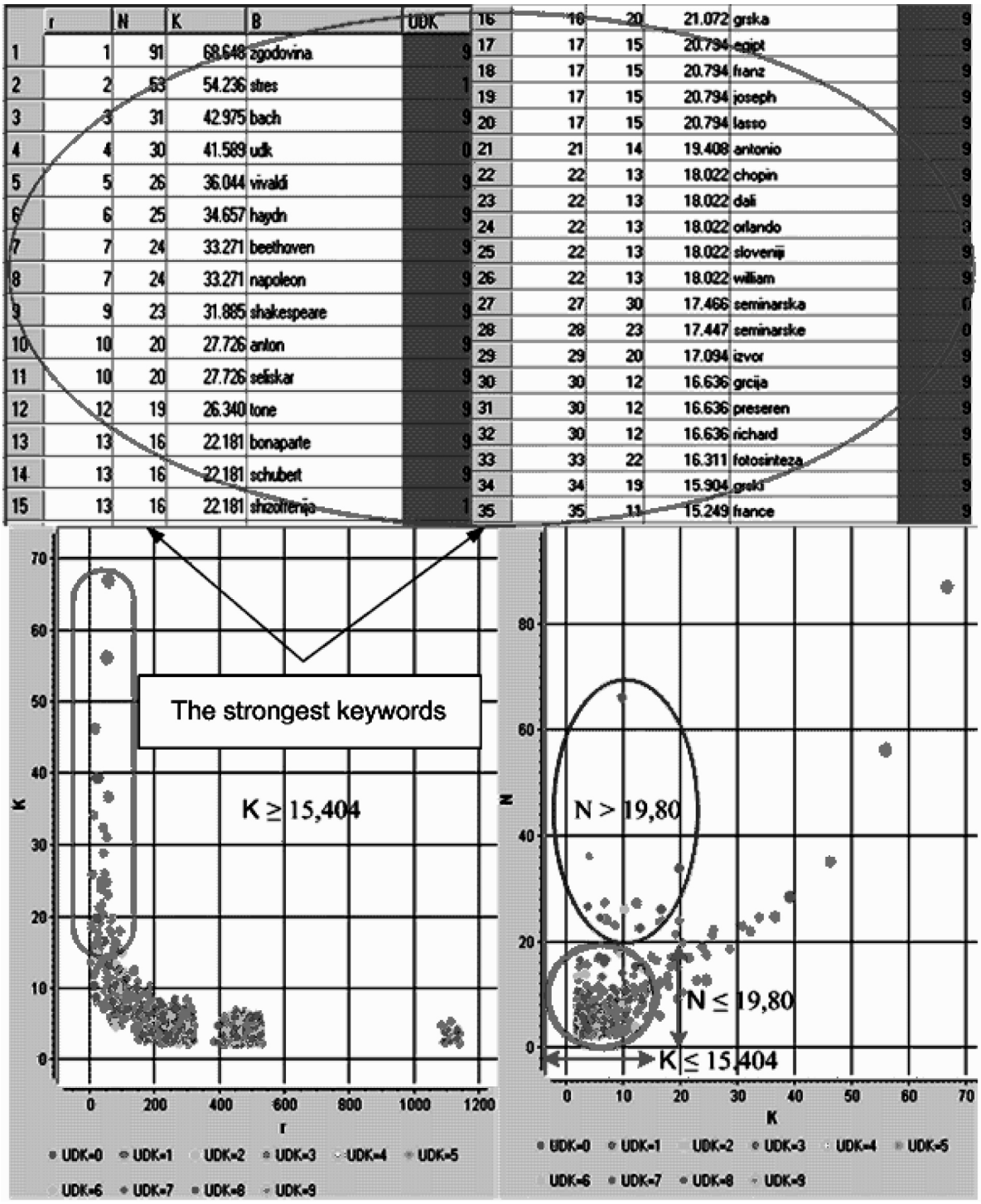

Figure 7. The strongest association rule on the basis of the keyness factor (K), rank (r), frequency (N) and UDC (UDK)

Cluster 2: UDC 6 and 1 (common denominators: existence of psychological and physiological healthiness, which means the precondition of life quality in societies, etc.)

Cluster 3: UDC 5 and 3 (common denominators: social conventions, traditions, clean social and natural environment, conditions of surviving, etc.)

Cluster 4: UDC 0 and 2 (common denominators: wide spectrum of organized knowledge, belief in gods and other supernatural things, etc.)

ID	∑K	K (%)	∑No	No (%)	UDC
1	272.25	5.05	5820	6.2	0
2	264.48	4.91	7479	7.97	1
3	208.16	3.86	9796	10.44	2
4	387.73	7.19	10910	11.63	3
5	499.53	9.27	10237	10.91	5
6	373.29	6.42	8174	8.71	6
7	318.7	5.91	11991	12.78	7
8	389.14	7.22	13830	14.74	8
9	2677.6	49.67	15573	16.6	9
Total	5390.9	100	93810	100	9 classes

Table 2. Comparison between the percentage of the most relevant keyness factors and user visits in the UDC classes

We compared all values (i.e., user visits on UDC pages, the UDC classified keyness factors of the relevant keywords and the frequencies of information retrievals on the internal search engine) with the intention to determine the ranks of the above mentioned clusters. The ranks were as follows:

Rank 1 was reached by cluster 1 (UDC 9, 8, and 7), the biggest cluster. These web users were identified as the most frequent and active.
Rank 2 was reached by cluster 3 (UDC 5 and 3). These web users were smaller in number and less active as those in the first cluster.
Rank 3 was reached by cluster 2 (UDC 6 and 1). The web users were less frequent and active than those in the third cluster.
Rank 4 was reached by cluster 4 (UDC 0 and 2), the smallest cluster. These web users were not very frequent and active.

Based on these measurements and findings the research questions (see subsection 1.3.2) could be answered as follows:

4. Our results confirm that it is possible to profile web users on the basis of the methods and tools described above.
5. The answer to the question of how many user profiles are to be distinguished is that it is appropriate to determine four different profiles of web users, which are ranked due to their information retrieval and page visit activities.

In the following subsection, we will profile and describe the characteristics of the web users more in depth.

3. The profiled web users

Each cluster we determined in the aforementioned analysis received a characteristic name: cluster 1 was called Adventurers, cluster 2 Observers, cluster 3 Applicables, and cluster 4 as Know-alls. For fast and easy comparison of the four user clusters, we created a table containing their basic characteristics.

Table 3 shows a comparison of the four user clusters (Adventures, Observers, Applicables, and Know-alls) and their basic characteristics (Activity of information retrievals and visits, UDC classes of activities, High ranked keywords and Extracted content of information retrievals). The distinguished user profiles and relevant details are shown in figure 8.

Figure 8 shows the segmentation of web users based on our measurements, analysis, insights, and distinguished basic psychological thought directions. The upper part of Figure 8 has already been displayed in figure 4, but, in the upper part of figure 8, an additional point of view was added, characterized as basic psychological thought directions. These are subsumed as food/drink desires, desires of success, health, love/loyalty, humor, comfort, fear, travelling and mobility, rivalry, harmony and cleanness, friendship, etc. (some directions like desires of construction, desires about new knowledge are not shown in figure 8). These thought directions are encompassed by the following global ways of thinking:

a. General or common way of thinking, which includes entities like feelings, food, holidays, sociability, fun, games, etc.
b. Philosophical way of thinking, which includes entities like science, art, business, innovations, etc.
c. Libidinal way of thinking, which includes entities like erotica, feelings, sentiment, comfort, etc.

In the lower part of figure 8, the global ways of thinking and the basic psychological thought directions are connected with packets from one to four, describing the closeness of UDC areas. The closeness of contents within the before mentioned directions and the UDC areas were evaluated in a similar fashion in the following two examples:

– food and drink desires are closely connected to the UDC area 6, where topics as food industry, food,

	Activity of information retrievals and visits	UDC classes of user activities	High ranked keywords (most frequently used by web users)	Extracted content of information retrievals
Adventurers	the greatest number of information retrievals and visits	9, 8, 7	history, Bach, biography, Haydn, Napoleon, Vivaldi, Egypt, Greek	geography, history of countries, biographies of famous people and families, important world history events, important works of literature and music
Observers	the second greatest frequency of information retrievals and visits	5, 3	photosynthesis, ecology, astronomy, mathematic, chemistry, genetics, osmosis, silver, iron, biology	natural sciences, social sciences, ecology, mathematics, physics, chemistry, biology, sociology, politics, conventions, science of law
Applicables	the third greatest frequency of information retrievals and visits	6, 1	stress, schizophrenia, satanism, India rubber	industrial medicine, commodity production, organisation of production, transport, engineering, manufacturing, philosophy, psychology
Know-alls	the smallest number of information retrievals and visits	0, 2	UDC, seminar, Hinduism, god, Islam	systematic of science, methodology, definitions, dictionaries, lexicons, catalogues, religions, bible, mythology, legends

Table 3. Profiled web users and their characteristics

etc. were classified. These elements are close to the general kind of thinking.
– love and loyalty are connected with UDC area 1. This example belongs to the libidinal kind of thinking.

Due to these crucial relationships and the aforementioned efforts, we could establish four different educational interests of web users (see figure 8: the central part of the big picture). These web user profiles were subsumed above as adventurers, observers, applicable, and know-alls and will be described in more detail in the paragraph below.

Adventurers – the most prominent group of web users, according to the highest number of information retrievals (5538 information retrievals) and the highest ranks of their keywords: e.g. history (keyness factor = 68.684), Bach (Keyness factor = 42,975), Haydn (Keyness factor = 34,657), Napoleon (Keyness factor = 33,271), Egypt (Keyness factor = 20,734), Greek (Keyness factor = 16,636), and biography (Keyness factor = 4,159). On the UDC 9, 8 and 7 webpages, we also noticed the greatest frequency of visits (41394 visits). Based on the web activities of the "Adventurers," we also identified connections between the UDC areas 9, 8, and 7 and the global interests of this user group, which are more strongly related to amusement, games, good feelings, food, holiday, sociability, art and, to a lesser extent to science/technology, innovative and business. From this, some potential connections to some basic psychological thought directions were derived [e.g., curiosity, pleasure to hunt, to play games, interested in sports, aesthetics/harmony, humor, ambitious (desires of success), and rivals]. These web users retrieved information from UDC areas 9, 8, and 7 (e.g., geography, history of countries, biographies of famous people and families, important world history events, important works of literature and music).

Observers - this profiled group of web users was the second most significant according to the number of information retrievals (3,119 information retrievals) and their keyword ranks: e.g. photosynthesis, ecology, astronomy, mathematic, chemistry, genetics, osmosis, silver, iron, and biology. They had used a relatively high number of keywords with a keyness factor

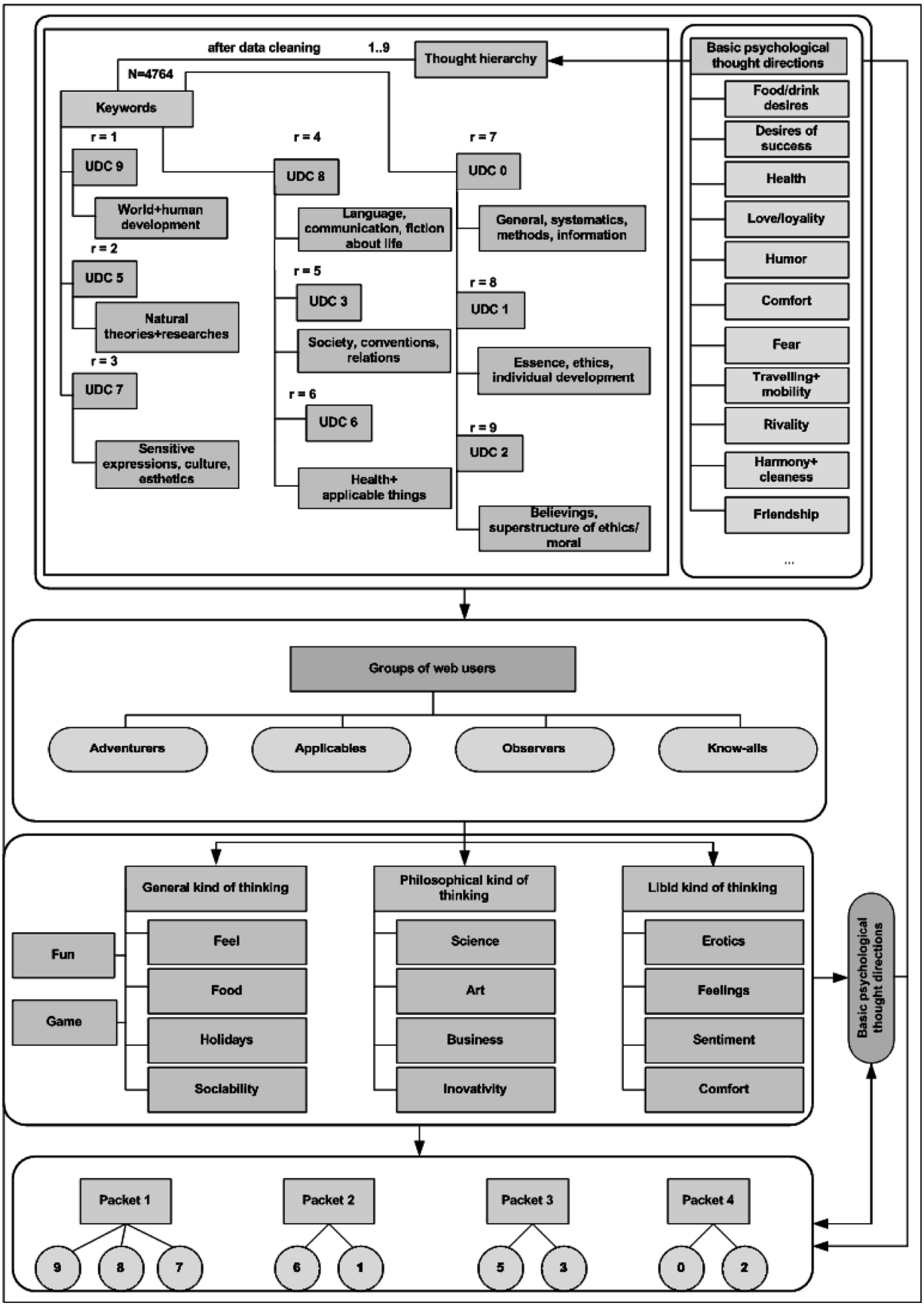

Figure 8. Segmentation of web users (user profiling)

ranging from 16.311 to 2.093. On the UDC 5 and 3 webpages, we noticed the second highest frequency of visits (21,147 visits). We identified some potential global ways of their thinking, which were in a stronger or weaker relationship to some basic psychological thought directions. These information retrievers were potentially more philosophical oriented and, to a lesser extent, general or libidinal. Their global interests contained topics strongly related to science, art, business, and innovation and less related to topics like games, feeling, food, holiday, sociability, and erotica. From this, we could derive potential connections to some basic psychological thought directions [e.g., aesthetics/harmony, creativeness, ambitiousness (desires of success), rivalry, loyalty, and cleanness]. These web users retrieved information from the UDC areas 5 and 3 (e.g., natural sciences, social sciences, ecology, mathematics, physics, chemistry, biology, sociology, politics, conventions, and science of law).

Applicables - this profiled group of web users reached the third rank according to our results. In comparison to the observers, the applicables created a lower number of information retrievals (2763 information retrievals), and some of their keywords reached very high ranks, e.g., stress (keyness factor = 54.236), schizophrenia (keyness factor = 22.181), Satanism (keyness factor = 13.863), India rubber (keyness factor = 12.477). On the UDC 6 and 1 websites, we registered the third highest frequency of visits (15,653 visits). As in the first two user profile groups, the applicables displayed some potential global ways of thinking that were in a more or less strong relationship to some basic psychological thought directions. These information retrievers were potentially equal to the philosophical, general, and libidinal ways of thinking. Their global interests contained topics that were more or less strongly related to science, art, business, innovation, games, feeling, food, holiday, emotions, etc. From these potential global thought characteristics, we could derive some potential connections to some basic psychological thought directions, which were, for example, curiosity, thirst for knowledge, creativeness, ambitiousness (desire of success), rivalry, loyalty and cleanness. These web users retrieved information from the UDC areas 6 and 1 (e.g., industrial medicine, commodity production, organization of production, transport, electrical engineering, machine engineering, civil engineering, manufacturing, philosophy of systems, philosophy of life, psychology, and stress).

Know-alls - this profiled group of web users reached the last or fourth rank, according to their information retrievals and visits on web pages UDC 0 and 1 they showed the lowest values (15616 visits and 1469 information retrievals). Some of their keywords reached surprisingly high ranks, e.g., UDC (keyness factor = 41.589), seminar (keyness factor = 17.466), Hinduism (keyness factor = 13.863), god (keyness factor = 12.540), and Islam (keyness factor = 10.971). As in the first three user profile groups, some of their potential global ways of thinking were in a more or less strong relationship to some basic psychological thought directions. These information retrievers were more philosophically oriented. Their global interests contained topics that were more strongly related to science, art, business, innovation, and less strongly related to games, feeling, food, holiday, emotions, sociability, etc. From these potential global thought characteristics, we could derive some potential connections to some basic psychological thought directions [e.g. thirst for knowledge, creativeness, ambitiousness (desire of success), rivalry, loyalty, and cleanness]. These web users retrieved information from the UDC areas 0 and 2 (e.g., systematic of science, methodology, definitions, dictionaries, lexicons, catalogues, religions, bible, mythology, and legends). These relations were relatively weak because these web users were mainly searching for short definitions in web lexicons and encyclopedias.

In the preceding review, we introduced the web user profiles distinguished according to their education interests. Almost all user activities (information retrievals and visits) were guided by high school educational programmes (e.g., information retrievals for preparing a seminar or investigation on topics like life and work of composers, writers, countries, osmosis, mechanics, religion communities, information science, stress in schools, and antic philosophy). The pilot web library, which was created in the years 2001 and 2002, was essentially an emulation of a classic high school library where the library items were classified according to UDC areas. After these procedures of web user profiling, we were able to construct the final idea of a new knowledge management solution called multidimensional rank thesaurus. The construction of a multidimensional rank thesaurus is the result of all previous efforts and, consequently, of the user-oriented approach in web libraries.

4.0 Multidimensional Rank Thesaurus

4.1 Multidimensional Rank Thesaurus and its Purpose

The MRT module (and its dimensions in particular) is adapted to the users of the education process in secondary schools facilitating their information retrievals on the basis of pre-prepared information, thus saving them a lot of time. The MRT module is the result of a multiple year study of both the education process and the user activities on the web. When searching for information, users could use an ordinary index search engine or a user thesaurus in which the keywords were ranked on the basis of previous information retrievals and user profiles. The thesaurus made it possible to include the users into other services of the MRT. The services range from participation in specialized discussion forums, web clubs, research activity, to the creation of personal information systems, etc. within the MRT. The user interests could be also monitored and allows constant adaptations of the MRT. The user MRT already includes important keywords which topped the users' past retrievals, as well as links to other terms being more or less closely related to the keywords. The user thesaurus offers a very simple and useful search function (Broughton 2006). It works on the principle of pre-prepared information so that a certain user profile obtains a maximum of useful information in one spot. Information is also classified by the UDC system, which additionally facilitates searching for information, in particular for librarians. The MRT organizes information on the basis of interest profiles of users, which show different relations between descriptors (*hierarchy, equivalence, association, synonym, antonym, homonym*).

The discussed MRT includes three different dimensions which are closely related, namely:

- Programme of the secondary school educational process
- Profiled users
- Ranked descriptors, which were the result of information retrievals on the internal search engine of the pilot web library.

For a better illustration, we are presenting the individual dimensions, the relations between them, and, in the final stage, a cross section of the MRT prototype (see figure 9).

Figure 9 presents the relations between the dimensions of the MRT:

- the dimension of keywords and descriptors, which achieved certain ranks and K factors (the K factor represents the importance or relevance of a particular keyword within the information retrievals) on the basis of measurements and analyses;
- the dimension of profiled users who often used certain keywords in their information retrievals during the studied periods;
- and, finally, the very important dimension of the school programme, which influenced many information retrievals.

In figure 9, it is noticeable that the first two dimensions are, to a great extent, dependent on the school program dimension, and the ranked descriptor dimension is, additionally, highly dependent on the profiled user dimension (characteristic keywords used in information retrievals). The keywords were the result of the school programme dimension to the extent that

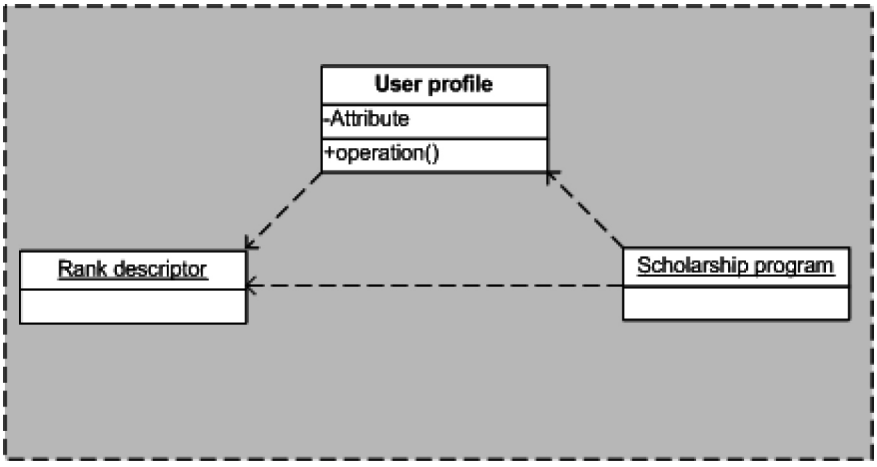

Figure 9. Dimensions of a multidimensional rank thesaurus (MRT)

the users were influenced under that dimension in their information retrievals. According to our estimate, roughly 90% of information retrievals were influenced by the collective school programme while 8-9% of information retrievals resulted from individual aspirations of the users. A nearly negligible percentage of retrievals could be categorized as noise, i.e., difficult to evaluate as either collectively or individually induced aspirations of users. The discussed MRT could conditionally be deemed a thesaurus containing and linking three different thesauruses in one. From the causal and condition-consequence point of view, we could say that the school programme dimension is superior to the user profile and ranked descriptor dimensions, while the user profile dimension is superior to the ranked descriptor dimension, where the view is not coherent from the global development aspect. Note that we have extrapolated the user profile dimension on the basis of measurements and analyses of information retrievals by the users in the pilot web library who were mostly searching under the influence of the collective school – programme. The dimensions are essentially in a sort of poly-hierarchic relationship in which they can exist either as equal (equivalence), relational (synonym and association), as antonym (inverse relationship – negative correlation), or hierarchical (superior : inferior). The variegated spectrum of relationships between the dimensions can change because of the users' target orientation (Mazzocchi and Tiberi 2009). The information focus of the users essentially determines the relations between the dimensions, because the solving of an information problem can essentially focus on one dimension (e.g., a student wants to write a research essay or to participate in a discussion forum), two dimensions (e.g., the teaching staff studies the students' interest in their teaching methods in order to improve its quality), or all three dimensions (e.g., the administrators of the MRT comprehensively study the system in order to improve the application and system).

Finally, let us present, in pictures and words, a cross section of the MRT prototype.

Figure 10 shows that users may select between the following options in the MRT:

- The course of psychology, a part of the secondary education curriculum (the school programme dimension).
- The profile of users subsumed as Applicables, subject to the user participating in discussion forums and/or communities for studies and/or exchange of experience and knowledge.
- The user decides to study the area of stress, which falls under psychology and is linked to other sciences and fields (e.g. medicine, neurology, psychiatry, sociology and physics), in more detail.

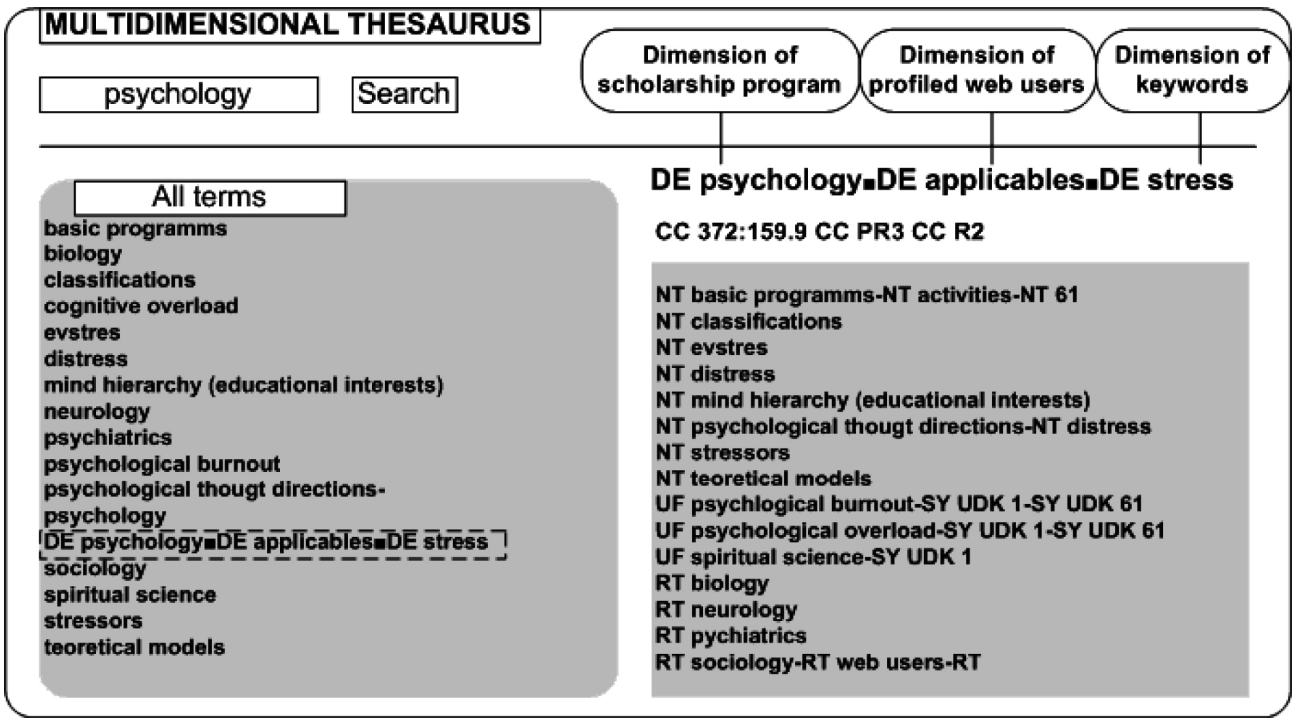

Figure 10. A fragment of the MRT prototype

– In the case of business users (e.g., teaching staff, librarians, archiving staff, IT staff, administrators) and perhaps certain other users (e.g., students and trainees), it may occur that two or all three dimensions are required. From that point of view, the information on top marked CC (Classification Column) is very important because data in the MRT are arranged by the UDC system and/or applicable dimensions. In the given case, the classification marking CC 372:159.9 CC PR3 CC R2 tells a business user the following in particular:
  a) Field "Psychology: applicable: Stress" is classified as a course in the education process with UDC classifier 372, which is, in this case, superior to the area of psychology with the classifier 159.9.
  b) Marking "CC PR3" marks users profiled as "APPLICABLES" who were ranked third of the four profiled user groups in terms of frequency and intensity of activity in the pilot web library (information retrievals, visits); and
  c) "CC R2" – the marking tells the user that the keyword "stress" was ranked the second of all keywords and, therefore, has the second biggest K factor meaning. The keyword is very important for the profiled applicable and probably also for other profiled users (adventurers, observers and know-alls).

CC gives the user the important information about how the keyword was classified, who was its most frequent user and how it was ranked, which also enables forecasting future interest of other users.

In order to adapt the MRT to the needs of the users, it is important that the information is located close together, for this facilitates information retrieval, as well as the advisory work within the reference process by the librarian, etc. (Arms 2000). We would like to conclude this subsection by subsuming the purpose of the MRT:

– The MRT provides the users with a more adapted and, consequently, more efficient access to useful information (*this applies to web-based libraries as well as other types of digital libraries*).
– The MRT provides lower level of noise in search results.
– The MRT can be used as a tool in resolving information problems in information and documentation centers as well as libraries (e.g., *as a part of the reference process between the librarian and the user*).
– The MRT could facilitate social contacts with other users, in particular with those of similar interests.
– The MRT functions as an excellent research platform for user research and, consequently, contributes to improved adaptability of information offer and services (e.g., library and educational services).

5.0 Conclusion

After years of collecting, observing and analysing the activities of web visitors on the private version of the pilot web library, we concluded that we were able to obtain new knowledge on the information needs of users and, consequently, gain better understanding of their mental worlds. The latter can be attributed to web measurements and analyses (in particular, the power law method) and the use of specific methods from the data mining area (e.g., profiled users on the basis of their educational interests). The final result of all efforts was the prototype MRT. The dimensions within the MRT represent the focus of individual users and the focused topics. The main idea behind all efforts was to focus on the thought hierarchy of users, which should, in our view, be the main orientation in the development of state-of-the-art IT applications and/or systems (e.g., knowledge management applications). A MRT (and its dimensions in particular) could be, in fact, a suitable knowledge management solution adapted to users of the education process in secondary schools, facilitating their search on the basis of pre-prepared information thus saving them a lot of time (Canales 2007).

References

Arms, William Y. 2000. *Digital libraries*. Cambridge (Mass.); London: The MIT Press.

Bak, Per. 1997. *How nature works: the science of self organized criticality*. Oxford: Oxford University Press.

Berry, Michael J. A. and Linoff, Gordon S.2004. *Data mining techniques: for marketing, sales, and customer relationship management.* 2nd ed. Indianapolis: Wiley Pub.

Broughton, Vanda. 2006. *Essential thesaurus construction*. London: Facet.

Canales, Alejandro et al. 2007. Adaptive and intelligent web based education system: towards an integral architecture and framework. *Expert systems with applications* 33n4: 1076-1089.

Chapman, Pete et al. 2000. *Crisp-DM 1.0: step-by-step data mining guide.* Copenhagen: NCR systems engineering. Available http://www.crisp-dm.org/CRISPWP-0800.pdf (link last viewed 29 April, 2009)

Kaushik, Avinash. 2007. *Web analytics: an hour a day.* Indianapolis: Sybex.

Konchady, Manu. 2006. *Text mining application programming.* Boston: Charles River Media.

Mayr, Philipp. 2004. *Entwicklung und Test einer logfile-basierten Metrik zur Analyse von Website Entries am Beispiel einer akademischen Universitäts-Website.* Berlin: Inst. Für Bibliothekswissenschaft.

Mazzocchi, Fulvio and Tiberi, Melissa. 2009. Knowledge organization in the philosophical domain: dealing with polysemy in thesaurus building. *Knowledge organization* 36: 103-12.

Mccue, Colleen. 2007. *Data mining and predictive analyses: intelligence gathering and crime analyses.* Oxford: Butterworth-Heinemann.

Mena, Jesus. 2003. *Investigative data mining for security and criminal detection.* Oxford: Buterworth-Heinemann

Pabarskaite, Zidrina and Raudys, Aistis. 2007. A process of knowledge discovery from web log data: Systematization and critical review. *Journal of intelligent information systems* 28: 79-104.