Virtual Web Contents (VWC) for Personalized Presentation

Imran Ghani
Faculty of Computer Science and Information Systems
Universiti Teknologi Malaysia, 81310, Johor, Malaysia
imransaieen@gmail.com

Seung Ryul Jeong
School of Business IT, Kookmin Univerity, Seongbuk-Gu, Seoul, 136-702, South Korea
srjeong@kookmin.ac.kr

M. Irfan Khan, Israr Ghani
Barani Institute of Information Technology, Rawalpindi, Pakistan
irfanpitafi@gmail.com, israrpatafi@gmail.com

***Abstract*— This paper aims to equip the next generation of Web users with the idea of personalized presentation of Web contents by substituting some of the real Web contents with the 'Virtual Web Contents' (VWC). In order to realize this idea, we introduce a novel concept of Personal Conceptual Dictionary (PCD). The PCD is an ontology-based dictionary of a collection of user-defined concepts that is utilized to create a new personalized layer of VWC residing between the real web contents and the end-user. Hence, the end-users rather than interacting with all the real contents, may also interact with some of the 'Virtual Web Contents' creating an exciting environment making their data visualization more personal and browsing more flexible. Our hypothesis is that by integrating VWC with real Web contents, the Web of future will not just work different but also provides more personalized and pleasant experience to its users. We present some interesting results obtained from the initial implementation of the VWC idea.**

***Keywords*— Personalized presentation, intelligent user interface, ontology, virtual web contents**

I. INTRODUCTION

In real life, each individual has or likes to have his/her own different concepts about information that exist in the environment. Generally, he/she uses these concepts in different ways at different times in a personal fashion. This phenomenon also applies to the users of digital contents – Websites. Based on our observation, it has been noted that some of the users of Web face a number of problems related to the contents (text, images, videos, etc) as discussed Section 3. Based on the discussed real-life observable facts in Section 3, we propose a unique approach called Virtual Web Contents (VWC). Figure 1(a) illustrates the traditional architecture of real Web, while Figure 1(b) depicts the newly proposed idea, wherein VWC layer has been integrated.

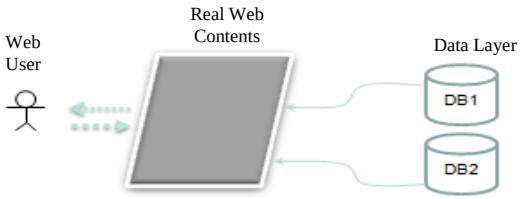

Figure 1(a): Traditional Web model

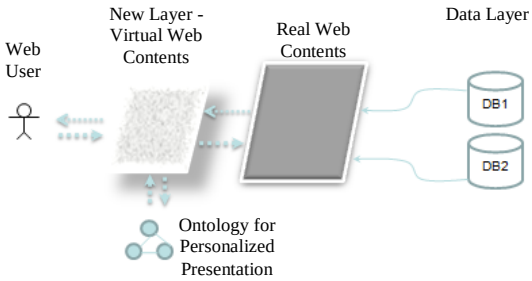

Figure 1(b):Virtual Web Contents (VWC) model

The basic idea behind the proposal of PCD is that each individual user can have his/her own personal dictionary where he can define new conceptual meanings to virtualize (substituting) the existing Web contents without changing the real contents. In the PCD model, a concept and keyword is a type of a metadata stored in Resource Description Framework (RDF) [17] format to be used for personalized presentation of information on user's screen. With this, an ontology-based user model is created containing each user's profile that integrates the PCD. Before going into the detailed discussion on real-life issues, it is appropriate to describe the motivation behind the idea of VWC.

II. MOTIVATION

Words are powerful. They relate to emotions and reactions. And like words, each person is unique. So, different words may have different meanings and feelings to different person. In real life, each individual has or likes to have his/her own different concepts about information that exist in the environment. Generally, he/she uses these concepts in different ways at different times in a personal fashion. This phenomenon also applies to the users of digital contents – Websites. And, it has been

noted that some of the users (including organizations such as educational institutes, parents as users) of Web face a number of problems related to the Web contents in terms of text, images, videos etc used on the Web. The detailed discussion on the issue is provided as follows.

III. FUNDAMENTAL ISSUES RELATED TO USER AND WEB CONTENTS

This research has observed a variety of issues and has categorized them as user centric (browser-based, personal) and lexical centric concerns.

A. User Centric Internet-based Issues

In the existing Web environment, the users have the following issues:

- The current browsers are not interactive and do not support mapping the user's personal concept about the information coming from Web servers.
- The users on client machine has no role to renovate (add, update) the Web contents (coming from server) according to their own meanings, thoughts or concepts that may make more sense of contents and easily understandable within their own context. In fact, any addition, removal, or change of content in a deliberate attempt to compromise the integrity of a Website is prohibited under Vandalism.
- The user has to read and understand Web contents written by many different authors that use several different terms for a single topic. Indeed, the user is lost in so many terminologies for a single topic.
- Certain indecent words (Filthy Words), the most notorious of which contain four-letters, are offensive to many people [5]. Many other people, of course, are not offended by these same words and may in fact frequently use these words themselves [3][4]. These very different reactions to indecent words as well as to other references to sexual or excretory functions make for potential conflicts and controversies. In this scenario, a user cannot change the real Web contents coming from Web server.
- There are constitutional laws [10] about obscenity (indecent exposure of words, actions, images) for mass media [4] and some of regulations have been implemented by several countries and organizations such as Communications Decency Act of 1996 (CDA) in USA, Obscene Publications Act in the UK, Obscenity and Indecency law in Canada [22], Internet Watch Foundation (IWF), Safety Net Foundation and so on [19]. However, due to the occurrences of obscenity, either no action is taken to prohibit publishing inappropriate Web contents, which shows that obscene publications laws are outdated or the entire Website is blocked using software filters (parents and school administrators also install such filters to avoid the harmful, violent, criminal activity-based, alcohol/tobacco inspiring, and sexual digital contents [20][23]). As a side-effect of taking this measure, the useful and harmless contents are also blocked.
- To date, there is no international law to regulate Internet obscenity and a number of Websites do not properly follow the local country's censorship policy. For instance, Wikipedia has this statement on its Website:

  "There are no forbidden words or expressions on Wikipedia, but certain expressions should be used with care, because they may introduce bias." [7]

- If each Website implements the rules for obscenity then the user need to login to multiple websites and enter their personal information and preferences, and the profiles are different for each site [2].
- The user has to find proper keywords that match with his/her thoughts and type the keyword every time he/she uses search engines.

B. User Centric Personal Issues

- Disabled people issues: Certain words create an emotional reaction to disabled people and found it really hard to read or see or hear [8]. This is so important that, some institutes online publish guidelines on appropriate language and behavior to use when interacting with disabled students [6].
  - The List of disability-related terms with negative connotations:
    - *"Cripple" used to mean "a person with a physical or mobility impairment."*
    - *"Diseased". Can be construed as an insult when referring to a non-transmittable disability.*
    - *"Dimwit". Used as an insult toward someone without a very high intelligence.*
    - *"Retard" literally means slow. It is used to describe someone with a learning disability*
    - *"Wheelchair-bound" for someone who uses a wheelchair*
- Age-biased language issues: Many women over the age of 30 find the term *"girl or miss"* offensive. On the other hand, saying *"ma'am"* to a lady under 40 or 50 would be offensive too. So, there is a problem in calling females/male at different age (this may happen during email communication).

- Religious/Ethnic/Racial Words: As instance, it is obligatory for Muslims to call the name of their prophet as Mohammad *(peace be upon him)*, not just Mohammad. In addition, in USA particularly, some African people feel uncomfortable to be called as "African American".
- Everyday words confusions: *expired/passed away, stupid, crazy or silly.*
- Other biased words: Word containing offensive or insulting sense/perception [9][11], have more than one meanings or extreme language [21].

C. Lexical Centric Issues

- Gender-biased language: In everyday life, the rules of grammar dictate that we use masculine pronouns (he, his, him, and himself) whenever a singular referent is required. In this case a doubt in Salutation is raised e.g., "Dear Sir/Madam".

In order to overcome the above issues, there is a need for a 'User-defined dictionary' that can enable the users to change and add the concepts into the dictionary which they like. It means this could be a good suggestion that a Web browser should provide the mechanism for a new type of user-defined dictionary that can be utilized for information access, visualization and representation. Indeed, this is what we are trying to do by introducing PCD. The PCD is user defined conceptual dictionary which is used to view the Web contents in a more personalized fashion. The PCD either can be integrated into the browse or portable or can be accessed via Internet.

IV. A NOTE ON RELATED WORK

The variety of research work is under way to create intelligent interfaces such as human-centered interfaces [16], intelligent interactive interfaces [12] [13], adaptive interfaces [15], personalized browser [2] personalized interfaces [1] [14] and so on. The work presented in this paper is under the personalized interfaces category. However, so far none of the existing approaches in personalized interfaces has attempted to personalize the Web contents as proposed in this article. Therefore, we suggest that there is a need to work more in the virtual Web contents area. The next section discusses the structure of PCD.

V. DESIGN OF PCD

The formation of PCD is not a usual dictionary or a thesaurus instead this is a component-based architecture. Its first component includes the attributes such as *Concepts, ConceptRules, keywordList* and *KeywordURI*. The second component contains user's concepts attributes. The third component is virtual visualization of Web contents. The main idea is to provide the user with a mechanism to replace the real content and view their preferred concepts (text or image) on the website. This is achieved by looking at the pair-wise relationships among concepts and key words. The PCD model maintains each user's PCD individually (Figure 2) which interacts with Intranet/Internet contents through semantic linking (knowledge routing) mechanism. This provides supports for personalized browsing, searching and visualization. The PCD model is designed to create a new GUI layer that facilitates two types of personalized operations: display virtual Web contents and help search user's concepts.

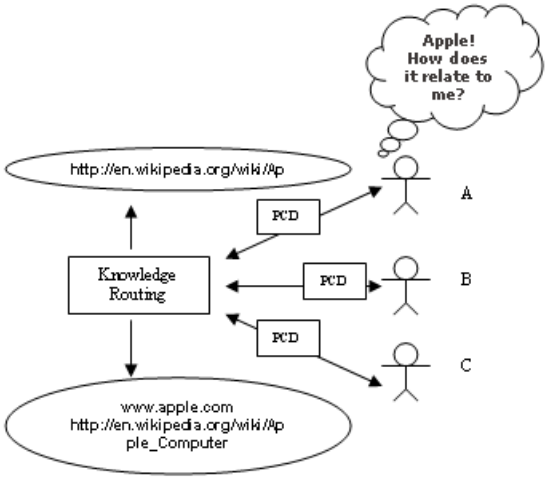

Figure 2: PCD implementation for individual users

The above Figure 2 illustrates the integration PCD and interaction between user and a real Web contents.

A. Structure of Concepts and Keywords

Similar to the SKOS impression [16] the notion of 'concept' in PCD is viewed as an idea; a unit of thought. However, what constitutes the PCD different is that it is a dictionary object that essentially deals with two types of words i.e. concept and keyword. with the n letters of each valid word combination used as the concept and keyword (used as tagging meta-data element), where n is the minimum number of characters that must appear in a word for it to be considered a word (we typically count two-letter words as valid). Typically a keyword consists of a root which is a concept. Each concept in the dictionary is a unique key that points to a list of keywords that start including any category of words i.e., single meaning, multi meaning or homographs, self-defined words and borrowed words from any other language to personal concept. Indeed, concept and keyword are list of strings that can be described as

alphanumeric list of letters constituting the concepts|concept1|concept2|...|concept N
alphanumeric list of letters constituting the keywords|keyword1|keyword2|...|keyword N

While keyword is a sorted list of four words in maximum where *ComapreTo()* used to sort the keywords alphabetically, *KeywordList()*.

Dictionary of triples (Object, Predicate, Subject) is stored in RDF format. Where
Object= concept, Predicate= containsKeyword, Subject= keyword

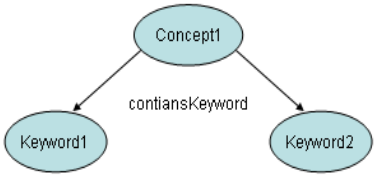

Figure 3: Hierarchal of concept-keywords

As shown in figure 3, PCD assumes that a dictionary is a list of entries in a hierarchical fashion. That is, the top level XML elements (concept as unique key) must contain all the dictionary entries as a list (where each element can have minimum one keyword value, its URI and concept rules such as to show picture, text/background color in case of text or show picture instead of text). This is a good storage form as shown in Figure below.

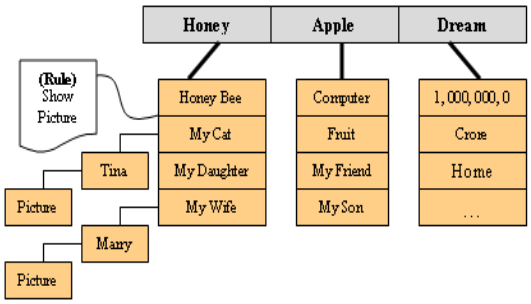

Figure 4: Keyword-Concept Ontological Structure

The structure in Figure 4 shows a subset of the dictionary keys (concepts) and their corresponding values (Keywords). It is used when viewing or searching for words. For instance it depicts three concepts: "Honey", "Apple", and "Dream". Each concept's value contains a list of strings that not necessarily start with the same value as the concept. For example, the key "Honey" has a list of strings as its value that includes words like "Honey Bee", “Tina”, "Maria", and “My Secretary”.

1) Concept and Keywords Structural Rules and Naming Conventions

In order to getting the dictionary into a suitable format that could help in easy manipulation of its XML/RDF data we defined some rules.

- Each concept in PCD can be any combination of characters from keyboard.
- In PCD, any word with two or more letters is valid; we typically count two-letter words as valid. As instance ‘Go’ is a valid word.
- So, the range of the concept should not be less than two characters and should not be more than 12 characters we check it by implementing maxWordsize() function.
- A concept is a unique word that can be linked with a keyword in the PCD. This is important to delimit concept-keyword linking otherwise it can be difficult to identify the related keyword for many occurrences of one single concept. With this semantics of words can be better treated if the user defines it according to its own concept. Getting generally accepted meanings without changing the personal concepts.
- We define the direction for the association of concept-keyword link as concept->keyword not vice-versa. The concept can only be unidirectional with keyword so only concept is displayed on the Web page not the keyword.

2) Functionalities Rules Hypothesis

We have defined the above rules to perform the following functionalities. The PCD is used for

- Concept selection and content visualization
- Defining rules to implement PCD for one site or global level (for all Web pages)
- Searching the concepts not the keywords
- Viewing the concepts: Search for ‘honey’, retrieve ‘Maria (my wife’s name)’ and display ‘honey’.

B. Attributes of Concepts

The second main component of PCD is the attributes list about their concepts. In fact, the user has been given flexibility to define the attributes he/she like to view on the Web pages, such as personalized text size, font, color (color blind or user with weak eye sight) and self-defined abbreviations for view management. The user’s concepts attributes list is stored in the Triple Store (XML/RDF) format so that it can further be used for semantic mapping of concepts with user’s preferences.

VI. SEMANTIC CONCEPT MAPPING

In practice, the PCD is simple but powerful implementation of the human concepts because the dictionary file is in a well-formed ontological structure. The PCD has two main engines: *Concept-Keyword-Contents Engine and Context-UserProfile Engine.* The working relationship between these two engines is a complex activity since they have to work together to produce the virtual Web contents on-the-fly. The reason is that because real Web concepts and user's preferred concepts are loosely coupled with adaptive behavior of the system. In order to semantically map (Figure 5) the concept with keywords a filtering mechanism (25) on RDF triples is applied with the Load function that starts by reading in the lines of the *dictionaryPath* XML/RDF file into the entries string array. It then iterates through each XML start and end tag element, which delimits each Triple (subject, predicate and object). These "pieces" of meta-data are loaded into the words string array to be further used by hash function to map the concept with user's preferences. A dictionary object is implemented that uses SPARQL Protocol and RDF Query Language (SPARQL) by implementing the functions:

pcd:equalMap ->implemented for the exact match just like '=' operator in SQL
pcd:likeMap -> implemented for the relative match such as 'like ' operator in SQL

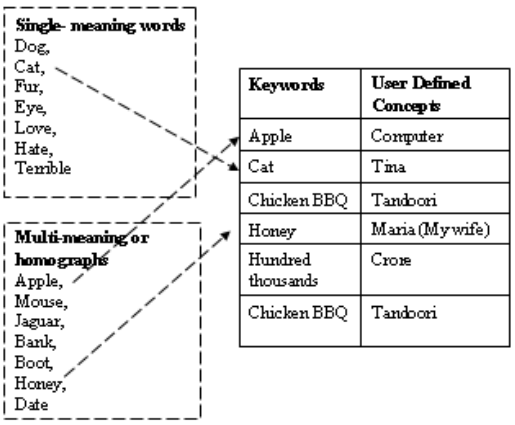

Figure 5: Concept Mapping

In this paper, we focus on nouns mapping only. However, in future, we will continue the semantic mapping of sentences and paragraphs including noun and verbs dealt by PCD.

VII. EXPERIMENTAL IMPLEMENTATION SCENARIO

Since this idea of virtualizating the Web contents is original and novel hence the scope of our implementation is not larger on global scale at this point of time. However, we have implemented two main functionalities of personalized presentation (adding concepts and keywords and visualizing them) and searching. We setup the LAN based client-server Website to implement the personalized system with PCD model. First we describe the communication flow between the user and PCD based system.

A. Contents Presentation Flow

The communication between the client's browser and the PCD is simple enough that a protocol based on HTTP and RDF/XML is sufficient. This standard defines several URLS that act as queries to the server. The HTTP processes on the server receives these queries, translates them to perform the query using SPARQL, and then parses the RDF/XML result, maps the concepts with users' PCD and send the virtualized Web pages back to the client's browser. The user browses the Web pages for his own purposes.

B. Adding Words by Selection

To populate PCD, a double click mechanism is used for a word that user likes to add (for virtual contents) while reading the contents (text in HTML document) of the Web page. The user simply selects a word by double clicking on it. Upon the double click the JavaScript event triggers the PCD. The PCD is prompted in a popup window (Figure 6). Now the user has three choices for the selected word.

(i) Add it as a new concept
(ii) Link it as a keyword with existing concept and define rules for it
(iii) Cancel the add operation.

The parent page is refreshed with newly defined contents. For example in Figure 6 the user added the keyword 'Madonna' under the concept of 'Honey'. As a result the contents of the page related to 'Madonna' are virtualized with 'Honey' without changing the real contents.

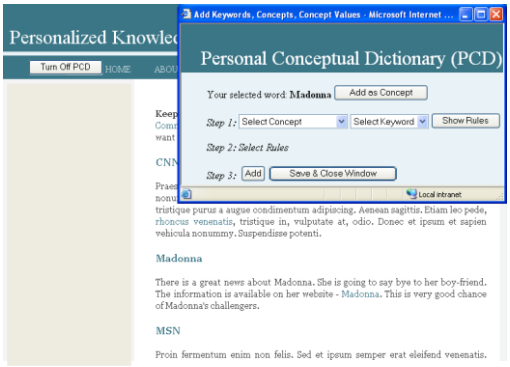

Figure 6: Adding Words by Selection

C. *Adding keyword/Concepts by Input*

The user has option to add keyword or concepts into the PCD since it accepts direct input from the user. This mechanism does not need to select the word from HTML page. The user can just click 'Add To PCD' to add any keyword or concept, if the same combination of letters does not exist already.

The PCD function getInput collects the user's input and solves the input validation by first creating a PCD instance based on the values entered by the user in the TextBox control (Figure 7) and then passes the data to addInput into relevant category of concept or keyword with rules. The addInput returns returns a true in case of successful insertion of record into RDF that contains the concepts for the keyword for that particular user. It return false in case of failure.

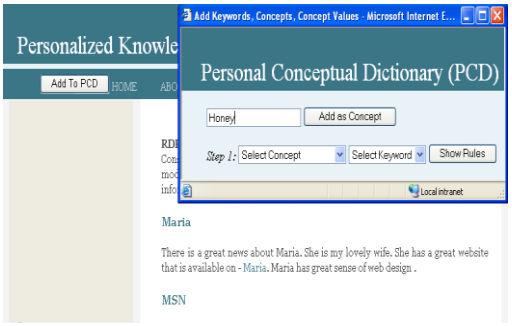

Figure 7: Adding Words by Input

In the above figure, user added 'Honey' as a concept for 'Maria' as a result the output Web page is virtualized with 'Honey' in the following Figure 8.

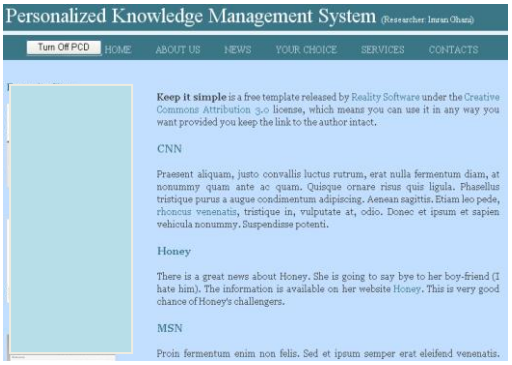

Figure 8: Virtualized Web Contents

C. *Searching by Concept*

In order to perform virtualized search, the user need to type the concept. The PCD translates that concept into the keyword and search contents for that 'concepts' from MYSQL database. Once PCD retrieves the contents, it virtualizes the 'keywords' into user-defined 'concept' and displays the contents related to the concepts values that user likes to see (Figure 9). However, the SafeSearch option has not been implement in our prototype since it already exists in Google SafeSearch (24).

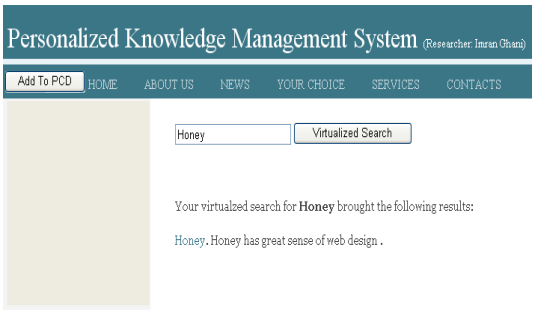

Figure 9: Virtualized Web Search and Results

VIII. APPLICATION AREAS

The concept of PCD can meet a variety of emerging needs for enhanced knowledge management, knowledge access, e-learning, and security and resolve all the issues mentioned in Section 2 of this paper.

- The users, making their web contents understandable, interesting, colorful and more personal for better experience and feelings.
- The users with special needs such as eye sight issues including color blind or weak eye sight or "I love colors" people.
- E-learning needs for instance if the color of a word is "brown" it means the origin of the word is Greece, *old Greek* word.
- Individuals or organization focusing on Safe Display (secure representation) of contents
- Team work terminologies
- Emotional Users
- Children story telling For Instance, once upon a time (show clock); there was a rabbit (show rabbit).
- Hiding unwanted Web contents from different users.

IX. FUTURE WORK AND CONCLUSION

We have presented our ongoing research work to virtualize the Web contents by ensuring personalization on a client-server setup. However, our next step is to work on the browser extension for Mozila Firefox with integration of PCD that will be managed on client's machine only where the contents of HTML pages will be vituralized on-the-fly according to the PCD rules settings. We argue that by integrating semantic-based VWC with the traditional Web will enable the next generation of Web users with more comfortable and exciting experience. This idea has the potential leading to

new integrative lines of research and development towards personalized presentation of Web contents.

ACKNOWLEDGEMENT

This research is supported by the Ministry of Higher Education Malaysia (MOHE) and collaboration with Research Management Center (RMC), UniversitiTeknologi Malaysia (UTM) under grant (Vot. No: 4D046).

REFERENCES

[1] Michal Tvarožek, Michal Barla, Mária Bieliková, "Personalized Presentation inWeb-Based Information Systems", OFSEM 2007, LNCS 4362, pp. 796–807, Springer-Verlag Berlin Heidelberg 2007.

[2] Melike Sah, Wendy Hall and David C. Roure "Designing a Personalized Semantic Web Browser", Proceedings of the 5th international conference on Adaptive Hypermedia and Adaptive Web-Based Systems, 2008.

[3]http://law2.umkc.edu/faculty/projects/ftrials/conlaw/filthywords.html [Retrieved: 18.05.2012].

[4]http://law2.umkc.edu/faculty/projects/ftrials/conlaw/indecentspeech.htm [Retrieved: 18.05.2012].

[5] http://en.wikipedia.org/wiki/Seven_dirty_ words [Retrieved: 20.05.2012].

[6]http://www.hw.ac.uk/welfare/disability-service/staff/disability-etiquette.htm [Retrieved: 20.05.2012].

[7]http://en.wikipedia.org/wiki/Wikipedia:Avoid_weasel_words#Unsupported_attributions [Retrieved: 17.05.2012].

[8]http://www.mtv.com/thinkmtv/features/discrimination/murderball/index3.jhtml#terminology [Retrieved: 17.05.2012].

[9] Online Slang Dictionary, http://onlineslangdictionary.com/lists/most-vulgar-words/ [Retrieved: 18.05.2012].

[10]http://caselaw.lp.findlaw.com/scripts/getcase.pl?court=us&vol=438&invol=726 [Retrieved: 16.05.2012].

[11] Peter Novobatzky and Ammon Shea, "Depraved and Insulting English", Publisher. Mariner Books, 2002.

[12] Marchionini, Gary; Geisler, Gary; and Brunk, Ben, "Agileviews: A Human-Centered Framework for Interfaces to Information Spaces", Proceedings of the ASIS Annual Meeting, v37 p271-80 2000.

[13] Francisco V. Cipolla Ficarra, "Advances in New Technologies, Interactive Interfaces and Communicability: Design, E-Commerce, E-Learning, E-Health, E-Tourism, Web 2.0 and Web 3.0", Proceeding of First International Conference, ADNTIIC, Huerta Grande, Argentina, October 20-22, 2010.

[14] Rennies, Jan, Stefan Goetze and Jens-E. Appell. "Personalized Acoustic Interfaces for Human-Computer Interaction." Human-Centered Design of E-Health Technologies: Concepts, Methods and Applications. IGI Global, 2011. 180-207. Web. 19 May. 2012.

[15] Les Miller, "Adapting Interfaces Based on User Needs", Proceedings of the Fifth International Conference on Advances in Computer-Human Interactions, 2012.

[16] Sudheendra Hangal, Abhinay Nagpal, Monica S. Lam, "Effective Browsing and Serendipitous Discovery with an Experience-Infused Browser", Proceedings of the 2012 International Conference on Intelligent User Interfaces. ACM, 2012.

[17] SKOS, http://www.w3.org/2004/02/skos [Retrieved: 18.08.2011].

[18]http://www.w3.org/2001/sw/Europe/reports/thes/rdfthes.html [Retrieved: 11.12.2011].

[19]Online Censorship: http://www.fourfootandclean.com/articles.htm [Retrieved: 18.05.2012].

[20] Samuel Slater1, "The commodification of violence on the internet: an analysis of 166 websites containing commodified violence", Internet Journal of Criminology, 2005.

[21] Traci Y. Craig and Kevin L. Blankenship, "Language and Persuasion: Linguistic Extremity Influences Message Processing and Behavioral Intentions", Journal of Language and Social Psychology, vol. 30, 3: pp. 290-310. 2011.

[22] Kirsten Krama, Richard Jochelson and GoverninG throuGh "Precaution to Protect equality and Freedom: obscenity and indecency law in canada after r. v. labaye [2005]", Canadian Journal of SoCiology/CahierS CanadienS de SoCiologie 36(4) 2011.

[23] Barracuda Networks, "Providing Safe Web Access in Educational Institutions", White paper, 2011. [Retrieved: 10.05.2012]

[24] SafeSearch, http://support.google.com/websearch/bin/answer.py?hl=en&answer=2521806&rd=1 [Retrieved: 12.05.2012]

[25] Imran Ghani, Choon Yeul Lee, Seung Ryul Jeong, Sung Hyun Juhn and Shafie A. Latiff, "A Role-Oriented Content-based Filtering Approach: Personalized Enterprise Architecture Management Perspective", (IJCSIS) International Journal of Computer Science and Information Security, Vol. 8, No. 7, October 2010.