

The Syllabus Based Web Content Extractor (SBWCE)

Saba Hilal
DAV Institute of Management
Haryana, India
saba21hilal@yahoo.com

S. A. M. Rizvi
Department of Computer Science
Jamia Millia Islamia
New Delhi, India
samsam_rizvi@yahoo.com

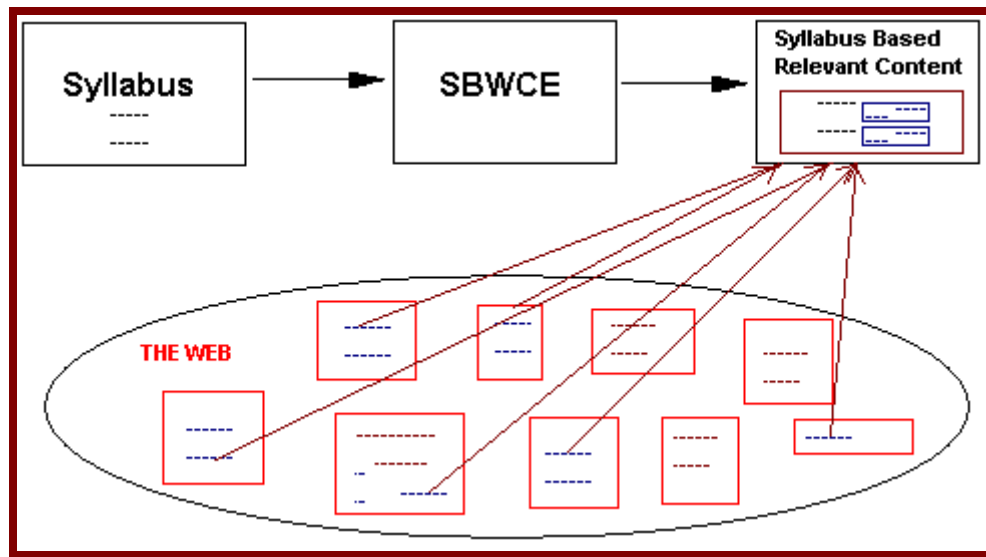
ABSTRACT

Syllabus Based Web Content Extractor (SBWCE) introduces a new technique of Syllabus Based Web Content Mining. It makes the Syllabus Based Web Content Extraction easy and creates an instant online book view based on the links relevant to the given Syllabus. Three important contributions are made by the current work. First, as multiple format educational information is needed for Syllabus based content; the technique used makes the finding of such content easier. Second, a new approach for capturing and recording the heuristics involved during searching by experts is used. Third, the grouping of Syllabus Words for precise extraction is exploited. This paper introduces SBWCE and presents the related details.

SYLLABUS BASED WEB CONTENT EXTRACTOR

According to Web-Based Education Commission -U.S. (2000), "The Internet is perhaps the most transformative technology in history, reshaping business, media, entertainment, and society in astonishing ways. But for all its power, it is just now being tapped to transform education". Still, Internet provides a great platform for e-learning. It includes Educational software, Programming languages, Educational Content Websites, School Websites, Virtual courses, Learning Management Systems, A-synchronous Learning Networks and Collaborative Learning Environments. According to ALN Report (Allen & Seaman, 2007), based on responses from over 2,500 colleges and universities, "Nearly 3.5 million students were taking at least one online course during the fall 2006 term and nearly twenty percent of all U.S. higher education students were taking at least one online course." It points to the strong need of the tailor made services and content designed around the needs of the individual and that, which is available at a time and place and in a form, which suits the learner's needs. The searching of educational information, content and material requires the development of better web content finding tools and techniques. Most of the times the required content is available on the Web but finding it is difficult. The Search Engines help to extract the information bundles from the vast ocean of the Web. However, finding of the correct collection still remains unsolved. Moreover, most of the time the search engine is not designed for the purpose that matches the user's search perspective. This also happens whenever the user searches for Syllabus Based Content. The need in such situation is to look towards content finding from a focused point of view. Web Mining is the application of Data Mining techniques to discover patterns from the Web (Etzioni, 1996). It can be effectively used for learning the on-line learner behaviour and for mining the content from the Web as per the demands of the learner. So, based on this motivation SBWCE is developed.

Figure 1: The Syllabus Based Web Content Extractor.



RELATED RESEARCHES

The development of SBWCE was based on the study of different important issues including the following-

Web Content Mining

Researchers are exploring ways to build systems that automatically gather and manipulate Web based information on user's behalf. But as the relevant content is embedded in HTML pages, extracting their content is difficult. A wrapper is a procedure for extracting a particular resource's content. Kushmerick, N., Weld, D., and Doorenbos, R. (1997), used Wrapper induction for information extraction and introduced wrapper induction, a technique for automatically constructing wrappers. Another related study by Crescenzi, V., Mecca, G. and Merialdo, P. (2001) describes a project RoadRunner to investigate techniques for extracting data from HTML sites through the use of automatically generated wrappers. Buttler, D., Liu, L., and Pu, C. (2001) presents a fully automated extraction system for the World Wide Web, called Omini. Omini parses web pages into tree structures and performs object extraction. Another important study by Chang, C-H., and Lui, S-L. (2001), describes information extraction based on pattern discovery. They present a technique that extracts information blocks without training examples using a data structure called a PAT tree. PAT trees allow the system to efficiently recognize repeated patterns in a semi-structured Web page. Rosenfeld, B., Feldman, R., and Aumann, Y. (2002) describe a general procedure for structural extraction, which allows for automatic extraction of entities from the document based on their visual characteristics and relative position in the document layout. Often fonts, physical positioning and other graphical characteristics are used to provide additional context to the information. This information is lost with pure text analysis. They also describe a specific implementation of the procedure to PDF documents, called PES (PDF Extraction System). Yu, Cai, D, S., Wen, J-R and Ma, W. (2003) proposes another study based on web content structure considering the visual representation. It specifies how a user understands web layout structure based on the visual perception and the scheme is independent of underlying documentation representation. The structural data extraction work by Arasu, A. and Garcia-Molina, H. (2003), focuses on web sites, containing large sets of pages generated using a common template or layout. They have studied the problem of automatically extracting the database values from such template-generated web pages without any learning examples or other similar human input. An important work by Gupta, S., Kaiser, G., Neistadt, D. and Grimm, P. (2003), uses DOM based Content Extraction of HTML Documents, to clean the web pages from unimportant images and extraneous links around the body of an article that distracts a user from actual content. They have implemented this approach in a publicly available Web proxy. Pinto, D., McCallum, A., Wei, X. and Bruce, W. (2003) focused on the ability to find tables and extract information from them. This paper presents the use of conditional random fields CRFs for table extraction and compares them with hidden Markov models -HMMs. The study by Wang, J., and Lochovsky, F.

(2003) focuses on scheme to help users query, extract and integrate data from web pages generated dynamically from databases, i.e., from the Hidden Web. They describe a system called, DeLa, which reconstructs part of a "hidden" back-end web database. It does this by sending queries through HTML forms, automatically generating regular expression wrappers to extract data objects from the result pages and restoring the retrieved data into a table. Chriisment, C., Dousset, B, Karouach, S, and Mothe, J. (2004) presented the work on extracting, exploring and visualizing geo-referenced information, presented a platform where the documents are analyzed in order to extract predefined elements such as location referenced. Then a multi-dimensional representation is extracted from each document and summarized in the form of tables from which the information is extracted. They combine mining and visualizing displaying the information graphically. Liu, B and Zhai, Y., NET (2005) presented automatic extraction of structured data from Web pages. Given a page, this method first builds a tag tree based on visual information. It then performs a post-order traversal of the tree and matches subtrees in the process using a tree edit distance method and visual information. It is based on finding data records and extracting items from them. The method works on both flat and nested data records. Zhai, Y., and Liu, B. (2005) studied the extraction of Web data using instance based learning. In contrast with the other approaches to data extraction including wrapper induction and automatic methods, they proposed an instance based learning method, which performed extraction by comparing each new instance to be extracted with labelled instances (or pages). The method works without an initial set of labelled pages to learn extraction rules as in wrapper induction and the algorithm is able to start extraction from a single labelled instance.

Different Approaches for Applying Searching Expertise

As, the development of SBWCE was based on application of expert search heuristics in form of Expert Filter Tokens; it was essential to study the different techniques to apply the search expertise. It was also about the studies that provided the tips for better content selection. The related studies like (Dodge, 2006), (Dartmouth.edu, 2005), (Manuel, 2003), (Pandia, 2007), and (The Web Support Team, 2007), presents the different ways used to enhance searching and getting better results.

Approaches for dealing with the unstructured Syllabus

Syllabus related researches are concerned with the recognition of a Syllabus as independent document, its classification, segmentation and analysis of the imposed structure. The important researches (Matsunaga et al., 2003), (Cohen, 2005), (de Larios-Heiman & Cracraft, 2006), (IDA et al., 2005), (Takasu, 2003) and (Xiaoyan et al., 2007), are included here.

The Web API Usage

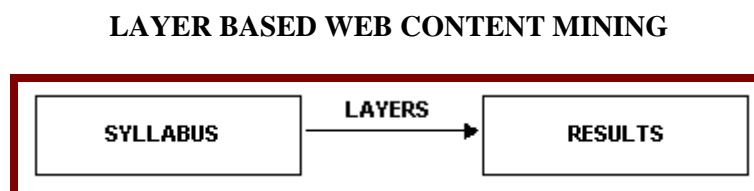
This included the study of some Information Retrieval Systems that used GoogleAPI such as (Brian D. Davison et al., 2003), (Dromey et. al., 2005), (Hsinchun et al., 2007), (O'Reilly & Dunnion, 2007) and also the study of Other Web APIs such as ProgrammableWeb.com, OpenSearch and Lucene Search Engine (Paul, 2004).

Different Approaches for achieving quality assurance related to Education

Several studies concerning attaining quality in various aspects of education have been dealt with but few very prominent and related studies such as (UNESCO) have been examined in order to get an overall picture of quality in and of education.

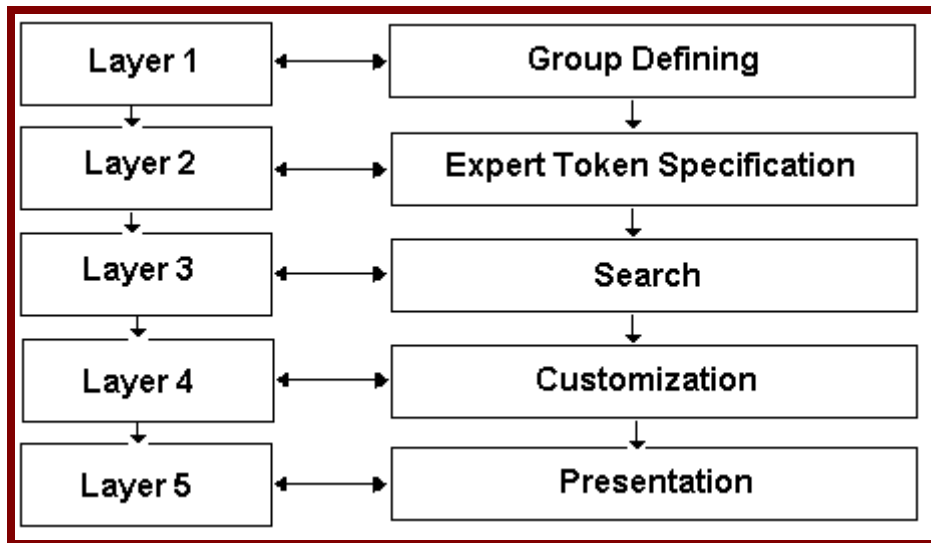
Our approach is based on using a layer based Web Content Mining Technique for Syllabus Based Web Content Extraction.

Figure 2: Layers between Syllabus and Output of SBWCE.



The different layers of SBWCE are given here.

Figure 3: Layers of Syllabus based Web Content Mining.



SBWCE DESIGN

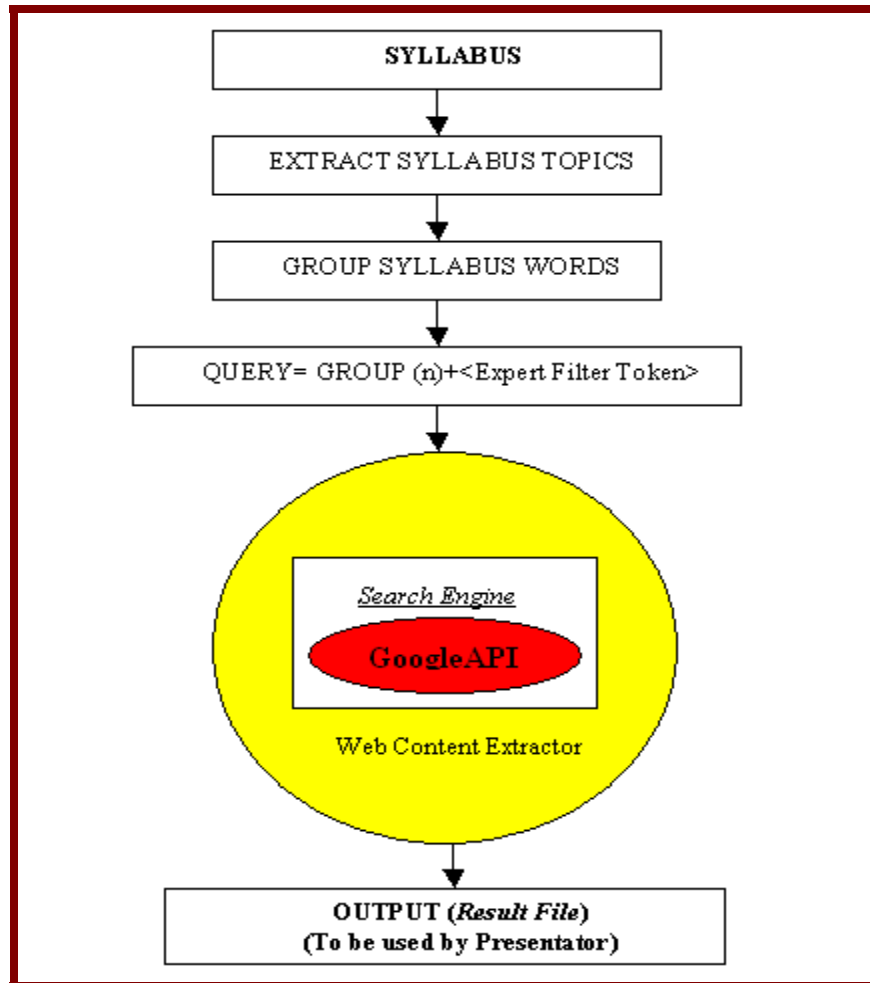
SBWCE is developed using Java. Its Top Level Menu is given below:

Figure 4: Top Level Menu for SBWCE.



The Google Interface

Figure 5: Search Engine Interface for SBWCE (Saba, 2007).



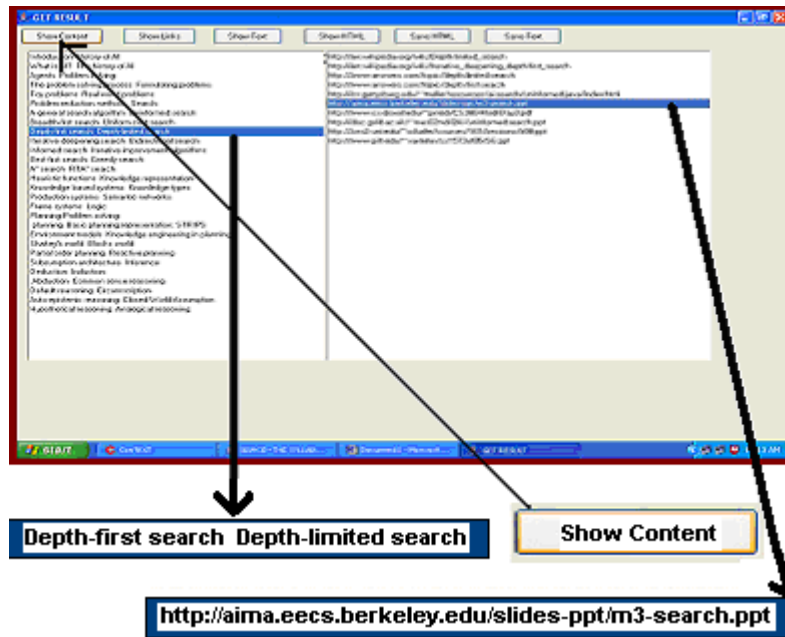
SBWCE RESULTS

The output screen provides a view similar to that of reading an online book where the user can go to a chapter and can view the subsections by following the links. However, merging all SBWCE's extracted Syllabus Related Links creates this instant book like view.

Here, the user first selects a link belonging to a Specific Topic Group and then selects the "Show Content" button in the top left corner for viewing the related Content.

The output of SBWCE can be used for mobile higher education systems, for uses ranging from creating health awareness to creation of instant user interfaces.

Figure 6: SBWCE Interface showing the extracted links for a Syllabus Topic Group.



USAGE AND IMPORTANT OF SBWCE

SBWCE can work for Syllabi of different types and for different content formats required by the user. It is useful for the different types of users, including kids. It can also be used for on demand creation of Syllabi and the related Web Content Extraction and also for user selected Syllabus topics. It covers different subject areas but uses English Language as the base language. The Syllabus Based Web Content Extractor is an important tool for researchers. It can be used for education / learning and business purposes. Automatic Syllabus Content Extraction is to help teachers / students in getting recent updated information even if the syllabus is an old one. It is useful for all those involved in educational content development process and can become an important part of future educational systems.

EVALUATION OF SBWCE

The evaluation of SBWCE was done in steps. The evaluation of the filter tokens was done through experimental case studies such as (Saba, 2007). Precision and Recall calculations were done for comparative analysis between SBWCE and Google subject results. For finding Y (The total relevant documents, on 1st page retrieved by Google) for a particular query Q) and D (The total relevant documents, extracted by SBWCE, for the same Q), the subject categories - General Subjects, Subjects difficult to consolidate, Subjects having too much Web Based Material and Critical Subjects that are not essentially Web Based, were used. The results were found to be more relevant if SBWCE was used.

CONCLUSIONS

Syllabus Based Web Content Extractor is developed for easy, efficient and effective Syllabus Based Web Content Mining. SBWCE is important because it is capable of satisfying the ever-increasing work expectation for accessing on demand learning materials. Life long learning can be made possible as SBWCE brings the concept of anytime and anywhere learning to reality. SBWCE is developed as a research product using JAVA. However, its development details such as coding and validation steps are not presented here. This paper has presented SBWCE as a tool for Syllabus Based Web Extraction and discussed the technical issues related to its development.

REFERENCES

- Allen, E. & Seaman, J. (2007). *Online Nation- Five Years of Growth in Online Learning*, Sloan C View. <http://www.aln.org/publications/survey/index.asp>.
- Arasu, A. & Garcia-Molina, H. (2003). *Extracting Structured Data from Web Pages*, SIGMOD-03.
- Bernie Dodge (2006). *Quiz on Searching Techniques*. <http://www.quia.com>,
- Buttler, D., Liu, L., & Pu, C. (2001). *A fully automated extraction system for the World Wide Web*, IEEE ICDCS-21.
- Chang, C-H. & Lui, S-L. (2001). *IEPAD: Information Extraction Based on Pattern Discovery*, WWW-10.
- Chriisment, C., Dousset, B, Karouach, S, & Mothe, J. (2004). *Information mining: extracting, exploring and visualising geo-referenced information*. ,SIGIR-04 Workshop on Geograpghic information retrieval.
- Crescenzi, V., Mecca, G. & Merialdo, P. (2001). *Roadrunner: Towards automatic data extraction from large web sites*. VLDB-01.
- Cohen, D. (2005). *About the Syllabus Finder*. Center for History and New Media. George Mason University. <http://chnm.gmu.edu/tools/syllabi/about.php>
- Dartmouth.edu. (2005). *Web Searching Tips and Techniques*. Last update 25-March-2005 by Biomedical Libraries Web Group, <http://www.dartmouth.edu/~biomed/workshops/search.html/>.
- Davison, B. D., Deschenes, D. G. & Lewanda, D. B. (2003). *Finding Relevant Website Queries*. Twelfth International World Wide Web Conference, pages 502-503, Budapest.
- de Larios-Heiman, L. & Cracraft, C. (2006). *SylViA: The Syllabus Viewer Application*. <http://groups.sims.berkeley.edu/sylvia/>
- Dromey, S., Heavin, C. & Neville, K. (2005). *Combining Website Search Engine Optimization With Advanced Web Log Analysis*, csrc.lse.ac.uk/asp/aspecis/20050112.pdf
- Etzioni, O. (1996). *The World-Wide Web: Quagmire or Goldmine?* *Communications of the ACM*, 39(11), 65-68.
- Gupta, S., Kaiser, G., Neistadt, D. & Grimm, P. (2003). *DOM based Content Extraction of HTML Documents*, WWW-03.
- Hilal, S. & Rizvi, S. A. M. (2007). *An Experimental Study to Identify the Impact of Expert Filter Tokens for Syllabus Based Searches*, presented in 18th Annual Conference {International Information Management Association}- Global Influences ~ The Networked Environment, University of Science and Technology, Beijing – China, (2007), paper published in refereed journal (CIIMA).
- Hilal, S. & Rizvi, S. A. M. (2007). *Using Expert Filter Tokens For Syllabus Based Web Content Extractor*, presented for ICIP 2007-International Conference on Information Processing, and published in proceedings of ICIP-2007.
- Hsinchun C., Xin L. , Chau, M., Yi-Jen, H. & Tseng, C. (2007). *Using Open Web APIs in Teaching Web Mining*, ai.arizona.edu/hchen/chencourse/webapi.pdf.
- IDA M., Nozawa Takayuki, Yoshikane Fuyuki & others (2005). *Development of Syllabus Database and its application to comparative analysis of curricula among majors in undergraduate education*. Research on Academic Degrees and University Evaluation.

- Kushmerick, N., Weld, D., & Doorenbos, R. (1997). *Wrapper induction for information extraction*, IJCAI-97.
- Liu, B. & Zhai, Y., NET (2005). *A System for Extracting Web Data from Flat and Nested Data Records*, WISE-05.
- Manuel, K. (2003). *Searching Techniques*. F:\Handouts\searchtechniques.doc, <http://lib.nmsu.edu/instruction/handouts/searchtechniques.PDF#search=%22searching%20techniques%22>, Last printed 07/21/2003 11:55 AM.
- Matsunaga, Y., S., Y., Ito, E. & Hirokawa, S. (2003), *A web syllabus crawler and its efficiency evaluation*, In: Proc. ISEE.
- O'Reilly, M. & Dunnion, J. (2007). *Crossword Clue Solver, Information Retrieval & Incident Analysis*, UCD School of Computer Science and Informatics, Retrieved: Apr 26, 2008. <http://csiweb.ucd.ie:8080/News/OpportunitiesDay/projs/irincident.html>
- Pandia (2007), *Pandia's 17 recommendations for Net searching*, <http://www.pandia.com/goalgetter/recommendations.html>
- Pinto, D., McCallum, A., Wei, X. & Bruce, W. (2003). *Table Extraction Using Conditional Random Fields*, SIGIR-03.
- ProgrammableWeb.com, *API Profile: Yahoo Search*, <http://www.programmableweb.com/api/Yahoo>
- Rosenfeld, B., Feldman, R., & Aumann, Y. (2002). *Structural extraction from visual layout of documents*. CIKM-02.
- Takasu, A. (2003). *Bibliographic attribute extraction from erroneous references based on a statistical model*. In: JCDL '03: Proceedings of the 3rd ACM/IEEE-CS joint conference on Digital Libraries. Washington, DC, USA, IEEE Computer Society. 49–60.
- The Web Support Team (2007). *Web Searching of books and slides*, The University of Edinburgh. <http://webhelp.ucs.ed.ac.uk/services/search/slides.html>, Computing Services.
- Paul, T. (2004). *The Lucene Search Engine*, Adding search to your applications. <http://www.javaranch.com/newsletter/200404/Lucene.html>
- UNESCO. <http://portal.unesco.org/education/en/>, Retrieved Feb 15, 2006.
- Wang, J., & Lochovsky, F.(2003). *Data extraction and label assignment for Web databases*. WWW-03.
- Web-Based Education Commission (U.S). (2000). *Power of the Internet for Learning-Final Report of Web-Based Education Commission*. <http://www.ed.gov/offices/AC/WBEC/FinalReport/Section1.pdf>
- Wikipedia, the free encyclopedia, *OpenSearch*, <http://en.wikipedia.org/wiki/OpenSearch>
- Yu, X., Tungare, M., Fan, W., Perez-Quinones, M., Fox, E. A., Cameron, William, & Cassel, L. (2007). *Using Automatic Metadata Extraction to Build a Structured Syllabus Repository*. Lecture Notes in Computer Science, Springer Berlin, 4822.
- Yu, Cai, D , S., Wen, J-R. & Ma, W-Y.(2003). *Extracting Content Structure for Web Pages based on Visual Representation*. Fifth Asia Pacific Web Conference (APWeb-03).
- Zhai, Y., & Liu, B. (2005), *Extracting Web Data Using Instance-Based Learning*, WISE-05.