Communications of the IIMA

An Experimental Study to Identify the Impact of Expert Filter Tokens for Syllabus Based Searches

S.A.M Rizvi
Department of Computer Science Jamia Millia Islamia

Saba Hilal
DAV Institute of Management

Follow this and additional works at: http://scholarworks.lib.csusb.edu/ciima

An Experimental Study to Identify the Impact of Expert Filter Tokens for Syllabus Based Searches

S.A.M. Rizvi

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

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

ABSTRACT

People search using search engines like Google. While specifying the query, some other words (filter tokens) are used with the topic being searched. These help in the filtering of the relevant and effective search results. SBWCE is being developed to make the Syllabus based Web Content Extraction easy and automatic. As SBWCE is based on using Expert Filter Tokens, there was a need to study their impact. Therefore, a framework to study the impact of these expert filter tokens was used on 'Computer Organization and Assembly Language', for experimental study. This paper presents the related details.

INTRODUCTION

The new trend in teaching / learning scenarios necessitate the use of audio-visual aids. With the ease of Internet availability, the vast ocean of the knowledge present on the web needs to be explored. This has helped to remove the constraint of non-availability of information and made the teaching / learning process easier but challenging. People now try to spend more time on Internet than in the library. This leads to the change in the teaching concept. Now, the recent, updated and correct information on any subject area is required from the Internet, thereby reducing the use of only the recommended books by the University. The teachers, especially those in higher learning environments have started accepting the challenge of this changed situation. The teaching today involves the mandatory requirement of syllabus based web content searching, selecting the relevant results and then presenting the content to the learners. There is also limited time available because of the transition to multitasking educational setups. So, for most Web based search needs people use search engines like Google.

A person becomes a search expert if the search process is continued for similar types of searches over a period of time T, where T is the threshold value. The choice of the filter tokens by such experts makes the search process effortless, easy and efficient. A new approach of capturing, specification, selection and prioritization of these Expert filter tokens, is brought forward, (S.A.M Rizvi and Saba Hilal , 2007), and exploited as a technique for filtering the web based educational content in the Syllabus Based Web Content Extractor, SBWCE. The validation plan of SBWCE is designed to ensure that SBWCE will support the users' needs efficiently and correctly. For SBWCE, this means that users will be able to get the information they need efficiently from the Internet. It should cover a range of users with different expectations, backgrounds and experience. The validation plan also includes an evaluation of SBWCE by students and faculty. This necessitates the study of its impact by measuring the quality of the web-based output.

This study includes the user requirement analysis for educational syllabus searching purposes, the experimental stage where the expert filter tokens are applied and case study is conducted. The results are then analyzed and user feedback is taken.

DIFFERENT APPROACHES FOR RESOURCE IDENTIFICATION AND GENERALIZATION

Resource identification is an important aspect of Web Content Mining. It is the process of retrieving the intended web documents. It is done by web search and meta-search engines, or by crawlers. The current methods not only focus on analyzing static web sites but can also deal with constantly changing web sites, such as news sites, as given by (Mendez-Torreblanca, A., Montes-y-Gomez M., and A. Lopez-Lopez, 2002). Choosing the appropriate search engine for resource identification is dependent on various factors, like it should be readily available, and is designed for fast and accurate retrieval of valuable information. Even if a popular and successful search engine is to be used, the different formats for inputting queries, the presentation formats of the retrieval results and quality of retrieved information are also analyzed. In the context of information retrieval, the performance is measured by speed of retrieval, precision, and recall. Various researchers have suggested to exploit measures for the importance of a webpage (such as authority and hub ranks) and use these measures to rank result lists. (Diligenti M., Coetzee F., Lawrence S., Giles C.L., Gori M, 2000) has introduced the concept of context graphs to represent typical paths leading to relevant webpages. For resource identification, either the capabilities of any popular search engine are exploited, or another search engine / crawler can be designed. Focused web crawlers can be another alternative. Compared to web search engines, focused crawlers obtain a much higher precision and return new pages, which are not yet indexed, see (Martin E., Hans-Peter K.and M. Schubert, 2004). . For Syllabus Based Web Content Extractor (SBWCE), new directions for Syllabus Based Web Content Mining are presented and published (S.A.M Rizvi and Saba Hilal, 2005), and a new technique of Syllabus Based Web Content Mining is developed and used. Our approach in Web Content Mining is to make the collective use of Syllabus Word Groups and Expert Tokens and implement it via a layering scheme.

SBWCE's layering scheme has the flexibility of using the existing capabilities of a search engine like Google for resource identification. Forming Syllabus Word Groups filters out the non-relevant results.

Generalization is another aspect of Web Content Mining. Most methods use data mining techniques for discovering associations, clusters, classification trees and rules. For instance, a method for detecting association rules that describe the content of a set of scientific online papers is given in (L.Singh, B. Chen, R. Haight and K. Scheuermann , 1999). Extracting companies data from the web and constructing classification trees for predicting the growth of the economic sectors is suggested in (Ghani et al , 2000). (Gelbukh, G. Sidorov, A. Guzman-Arenas ,1999) proposes a method to categorize documents based on a weighted topic hierarchy. Clustering is grouping similar documents together to speed up information retrieval. Clustering web sites by their content is given in (Crimmins, A. F. Smeaton, T.Dkaki and J. Mothe, 1999). (Alexandrov, A. Gelbukh, and P. Makagonov, 2000) describe a method for clustering and classifying interdisciplinary documents based on qualitative and quantitative properties. In SBWCE, the use of Expert Filter Tokens during the query string generation for resource identification, leads to the automatic clustering of user preference based results.

A FRAMEWORK TO STUDY THE IMPACT OF EXPERT FILTER TOKENS FOR SYLLABUS BASED SEARCHES

While searching, the human point of view, decisions and tricks are recorded and filter tokens are specified. A user becomes a search expert if same types of searches are performed continuously, for more than time T. The promising result (filter tokens), are then given for review to peers involved in searching different subject areas and their opinion is taken. A questionnaire is designed to decide the priority and selection of the filter tokens. After this prioritize the Expert tokens to be used in searches. Broadly, for Syllabus Based Web Content Searching, the following steps are required.
1. Syllabus Specification
2. Analysis of the profile for the user who has to use the syllabus based web content.
3. Analysis of the web content requirements.
4. Searching the Syllabus Based Web Content using the final list of filter tokens.
5. Presenting the results to the user.
6. Taking the user's feedback.
7. Studying the impact and quality of the results achieved.

STEPS APPLIED FOR EXPERIMENTAL STUDY

Using the Syllabus of Computer Organization and Assembly Language (COAL)

Experimental Stage

Let T be the time required for capturing filter tokens. Here, T > 1 year.
Search Expert: Subject Faculty (*Author*)

Filter Tokens List used for prioritization:

Filter tokens were selected by continuing the search process of Syllabus Based Material for different subjects by the Search Expert, for duration of more than 1 year. The following list of filter tokens was then finalized and put forward for peer review.

[Slides, Notes, Lectures, Tutorials, Presentations, Diagrams, Formulae, Examples, Notes, Basics, Introduction].

The peer group was of those involved in the similar searches within the audio-visual teaching and learning setup. The following questions were given through a Search Questionnaire that was tailored using Reference 3: Resource Discovery Working Group, (2003), to know the opinion of peer group:

_____

Why are you using existing search engines (like Google) for your syllabus based educational work?
- To locate specific documents (e.g, articles or documents or dictionaries)
- To look for specific data such as
  - Slides
  - Notes
  - Lectures
  - Tutorials
  - Presentations
  - Diagrams
  - Formulae
  - Examples
- To find answer to syllabus based questions (e.g., what's _____?)
- To find people working in the same field / in the same area.
- To find methods and services for analyzing syllabus based resources.

What kind of search input types do you like to use?

- Simple keyword (Google-type) search
- Form-based search with predefined fields (e.g. language names, geographical areas, topics, subtopics, properties of resources)
- Natural-language questions

In the searches you perform, what kind of search results do you find helpful?

- Google style: Some data from the document displayed and links to the document.

- Subscription abstract service style: Title and abstract, the ability to check off particular search results and save them for later examination in some way.
- Laboratory style: The ability to choose and specify methods, to evaluate the results, to subject them to further available methods and to store the final results.
- Images and Diagrams
- Material that can be downloaded as PowerPoint slides, PDF documents or word documents.

Of the following sophisticated kinds of searches, what sorts do you think you would find most helpful?

- Find examples / description of some topics related to syllabus.
- Finding syllabus based material from educational sites.
- Finding syllabus based material from free e-books.
- Finding syllabus based educational content belonging to famous authors in a specific course area.

While answering these questions, if you were thinking of the any subject area, kindly specify.

Submitter Information:
Name:
Institute:
Designation:
Qualification:
Date:

_____

Analysis of Peer Group Opinions and Search patterns:

In response to the above questionnaire that was given to 34 persons, 73% people agreed that they use search engine (Google) for their syllabus based educational work, to look for specific data such as Slides, Notes, Presentations etc. The following list was given as the assessment of user priority.

[Slides, Notes, Lectures, Tutorials, Presentations, Diagrams, Formulae, Examples, Notes, Basics, Introduction].
The arranged list, in decreasing order of peer group priority response, including weights is given below:

[Notes (w=16), Slides (w=14), Presentations (w=13), Diagrams (w=11), Examples (w=11), Tutorials (w=7), Lectures (w=5), Formulae (w=1)]
85% of the respondents preferred simple keyword (Google-type) searches. 79% found useful, the material that can be downloaded as PowerPoint slides, PDF documents or word documents. 50% agreed that they like Google style results and 41% agreed that they require images and diagrams in their web based Content. 55% respondents were of the same opinion that examples / description of some topics related to syllabus were most helpful in the sophisticated kinds of searches performed.

STEPS FOLLOWED FOR " COMPUTER ORGANIZATION AND ASSEMBLY LANGUAGE"

Syllabus Specification

The Syllabus for Computer Organization And Assembly Language that was used for this experiment is given below:
_____

Representation of Information:

Number Systems, Integer and floating-point representation, Character Codes (ASCII, EBCDIC), Error detection and correction codes.

Basic Building Blocks:
Boolean Algebra, Flip-flops: RS Latches, D, JK, T and Master-slave, Registers, Buffer, Shift and Controlled shift registers, counters: Ripple. Synchronous and Ring Counters, Half adders and Full adders.
CPU Organization:
Control Unit Design, Micro-operations, Micro programmed vs. Hardwired Control Unit Implementation, Design of ALU, Peripheral Devices I/O devices (Video Terminals and Printers) and Controllers, I/O Techniques : Programmed and DMA, Storage Devices (Tape and Disks), Memory Hierarchy, Interleaved Memories, Associative memories.
Assembly Language Programming:
Programmers model of a machine, Overview of 8 to 32 bit processors. Assembly Language Programming with 8086/8088: Registers, Addressing modes, Instruction set, development of programs.

_____

Analysis of the profile for the user who has to use the syllabus based web content.

The Web based educational material was required for teaching purpose by the faculty. As there was no book available, having the complete syllabus content, the notes were needed to be distributed to the students. The students' details are given below:

Web Content Users: 60 MCA students, Year 2006
Course: MCA
Class: MCA First Semester
Institute: DAV Institute of Management, NH3, NIT, Faridabad, Haryana, India.

Analysis of the web content requirements

The content of this syllabus was required for teaching purpose. As there was no single book available in the library for this complete syllabus content, the desired format for the content retrieved was desired to be suitable for teaching purpose and needed to include basics, description, related figures and diagrams for each topic separately. The content in this case was required for MCA students and not for kids, so the issue of description in easy language was given less priority.

Most desired format for the Web Content: PowerPoint slides as it was the mandatory requirement for the audio-visual lecture.

Searching the Syllabus Based Web Content using the final list of filter tokens.

Time given for content retrieval:
Maximum 40 hours (One hour daily given for preparing the lecture of one hour duration)
The heuristic used in the search process selected one or the combination of the following "filter tokens" to be added with the topic being searched.
The query string that is generated is of the form
<Topic> <filter token>
For example, if the content is being searched for the topic "Interleaved Memories" and the filter token selected is "slides" then the query string will be
Interleaved Memories Slides
More than one filter token can also be selected and added with the query. If the filter tokens selected in the above example are "Lecture" and "Slides", the query string will be
Interleaved Memories Lecture Slides
Observations: The following points were observed during the syllabus searching process.
- Even if a single search engine is used, for example "GOOGLE", the addition of customized filter tokens can lead to effective search results in less time.
- The results obtained were capable of satisfying all the given objectives as specified by the users.
- The syllabus based retrieved content was capable of easy restructuring and reformatting and this made the tailoring of the lecture content, quick and easy.

Presenting the results to the user.

Improved teaching methodology was used using audio–visual classrooms and 400 slides (tailored) were given to MCA students, containing the syllabus content for the above given course.

Taking the users' feedback.

A questionnaire, in the form of user feedback form was given to the students after the course was complete. See figure 1 given below.

Figure 1: Feedback Form.

Student Feedback Form

MCA 1ˢᵗ Semester (Year 2006)

DAV Institute of Management, Faridabad, Haryana, India.

SUBJECT: Computer Organization and Assembly Language

Write Yes/No for each of the following:

1. The course material required for this course is available completely in one book. If Yes, write the name of the book._____.

2. The course material required for this course is available completely on a website. If Yes, write the address of the website._____.

3. All the topics of the syllabus were taught.

4. The PowerPoint slides were used in 95-100% of lectures.

5. The PowerPoint slides contain content from the Internet.

6. Do you know the "Search Engine" that was used for these lecture slides?

7. The PowerPoint slides contain enough no. of figures and diagrams.

8. Is this PowerPoint content reliable?

9. Is this PowerPoint content from educational web sources?

10. The complete syllabus content is covered using 360-400 PowerPoint slides?

11. Is this Internet based PowerPoint content useful for you?

12. Is this Internet based PowerPoint content current, fresh and updated?

13. The syllabus was completed in 35- 40 hours.

14. The syllabus content as compared to that available in the recommended book is

    a. Better

    b. not good

    c. same

Name / Signatures of the student:

Date:

Feedback Analysis

44 students were present at the time of feedback. Their feedback responses are given below:
100% students agreed that the course material required for this course was neither available completely in one book, nor it was available completely on a website.
95% students agreed that all the topics of the syllabus were taught and the PowerPoint slides were used in 95-100% of lectures.

81% of students were sure that the PowerPoint slides given to them contain content from the Internet and 98% students responded that they were not aware of the "Search Engine" that was used for getting these lecture slides.
88% students were satisfied that the PowerPoint slides contained enough number of figures and diagrams.

97% students agreed that the PowerPoint content was of educational web sources.

93% students agreed that the complete syllabus content was covered using 360-400 PowerPoint slides and also that the syllabus was completed in 35- 40 hours.
84% students responded that the given Internet based PowerPoint content was useful for them and 81% felt that the PowerPoint content was current, fresh and updated.

But, only 65% students were confident that this Internet based PowerPoint content was reliable and good enough as compared to that available in the recommended books.

## Studying the impact and quality of the results achieved

Keeping the positive feedback response threshold value greater than 80%, the feedback received was good in all points accept that the students were not very confident about the reliability and the quality of the PowerPoint content, as it was retrieved from the Internet. So, the next step was to wait for the students' final University results, to assess the quality of the Web based content and the usefulness of Expert filter tokens.

The following figure shows the MCA student's final result percentages and their comparison with the results achieved for the paper of Computer Organization and Assembly Language (COAL). The total result of the students, given below, is based on the theory and practical subjects including

1. Mathematical Foundation of Computer Science
2. IT and C Language
3. Computer Organization and Assembly Language
4. PC Software
5. System Analysis and Design

The good results of Computer Organization and Assembly Language clearly show the positive impact of the Web based Syllabus Content that was searched and selected using the Expert Filter Tokens and was distributed to the students. The impact of the teaching skill, in the said subjects was ignored as a factor, because the faculty had taught this subject for the first time during the last five years. This leads to the assessment that the Web based Syllabus Content was of good quality and the use of Expert Filter Tokens was a successful method.

Figure 2: Result Comparison (y axis showing marks %age).
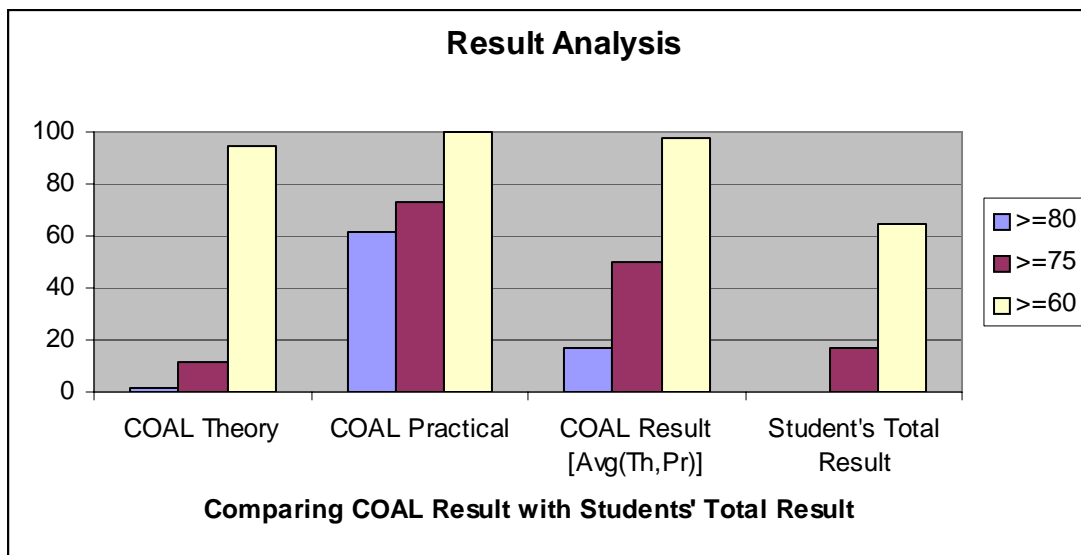
CONCLUSION

This paper presents the different approaches used for Resource Identification and Generalization and the experimental details of using the expert filter tokens for filtering the relevant Syllabus based Educational Content. The syllabus of Computer Organization and Assembly Language, required for MCA students, was used for experimental study. As there was no single book or recommended website having the complete course material

needed for the completion of the course, the required teaching and learning material was searched and extracted from multiple websites. While applying the expert filter tokens during search, the requirements for teaching and learning, desired content format, time limit given for Content Searching and the user's profile were taken care of. This experiment was conducted as a stepwise process, based on a predefined framework. The results showed the success of Expert Filter Tokens. This framework can also be used for the other subjects belonging to different categories.

REFERENCES

Alexandrov, A. Gelbukh, and P. Makagonov. (2000). Evaluation of thematic structure of multidisciplinary documents. Proc. NLIS-2000, 2nd International Workshop on Natural Language and Information Systems, IEEE Computer Society Press.

Crimmins, A. F. Smeaton, T.Dkaki and J. Mothe.(1999). TetraFusion: information discovery on the Internet. *Journal of IEEEExpert,* pp 55-62.

Diligenti M., Coetzee F., Lawrence S., Giles C.L., Gori M (2000). Focused crawling using context graphs. Proceedings VLDB 2000.

Gelbukh, G. Sidorov, A. Guzman-Arenas (1999). Use of a weighted topic hierarchy for document classification. Proceedings of the Second International Workshop on Text, Speech and Dialogue. Pages: 133 – 138.

Ghani   et al (2000). Data mining on symbolic knowledge extracted from the web. Proceedings of the Sixth International Conference on Knowledge Discovery and Data Mining (KDD-2000), Workshop on Text Mining. Boston, pp 29-36.

L.Singh, B. Chen, R. Haight and K. Scheuermann (1999). An algorithm for constrained association rule mining in semi-structured data. Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD 99), pp 148-158r, China, 1999.

Martin E., Hans-Peter K.and M. Schubert (2004). Accurate and efficient crawling for relevant websites, *VLDB* 2004. 396-407.

Mendez-Torreblanca, A., Montes-y-Gomez M., and A. Lopez-Lopez (2002). A trend discovery system for dynamic web content mining. Proceedings of the 11th International Conference on Computing. Mexico City, Mexico.

Resource Discovery Working Group (2003). Search questionnaire. EMELD Language Digitization Project Conference 2003.

S.A.M. Rizvi, Saba Hilal (2005). New directions in syllabus based web content mining. Souvenir of NSIF- 2007, JMI, Delhi. Also presented at National Conference on Emerging Technologies and Applications, ETA-2005, Rajkot, Gujrat, India, 2005.

S.A.M Rizvi, Saba Hilal (2007). Specification, selection and use of prerequisite filter tokens in searching of syllabus based educational content. Workshop on "Information Technology in Governance, Industry, Medicine & Environment", Dr M. C. Saxena College of Engineering & Technology, Lucknow.

S.A.M Rizvi, Saba Hilal (2007). Using expert filter tokens for syllabus based web content extractor. Selected for ICIP 2007, International Conference on Information Processing.