October 2011

Focused Crawling for Educational Materials from the Web

K.R. Premalatha

Deptt. of Computer Science & Engineering, College of Engineering, Guindy, Anna University, Chennai, India., premlatha.k.r@gmail.com

Follow this and additional works at: <https://www.interscience.in/ijcsi>

Part of the [Computer Engineering Commons](#), [Information Security Commons](#), and the [Systems and Communications Commons](#)

Recommended Citation

Premalatha, K.R. (2011) "Focused Crawling for Educational Materials from the Web," *International Journal of Computer Science and Informatics*: Vol. 1 : Iss. 2 , Article 7.

DOI: 10.47893/IJCSI.2011.1020

Available at: <https://www.interscience.in/ijcsi/vol1/iss2/7>

This Article is brought to you for free and open access by the Interscience Journals at Interscience Research Network. It has been accepted for inclusion in International Journal of Computer Science and Informatics by an authorized editor of Interscience Research Network. For more information, please contact sritampatnaik@gmail.com.

Focused Crawling for Educational Materials from the Web

K.R.Premalatha & T.V.Geetha

Deptt. of Computer Science & Engineering, College of Engineering, Guindy, Anna University, Chennai, India.

E-mail : premlatha.k.r@gmail.com

Abstract - An enormous amount of learning material is needed for the e-learning content management system to be effective. This has led to the difficulty of locating suitable learning materials for a particular learning topic, creating the need for automatic exploration of good content within the learning context. In this paper, we aim to tackle this need by proposing a novel approach to find out good materials from www for eLearning content management system. This work presents domain ontology concepts based query method for searching documents from web and proposes concept and term based ranking system for obtaining the ranked seed documents which is then used by a concept-focused crawling system. The set of crawled documents so obtained would be obtained an appropriate set of content material for building an e-learning content management system.

Keywords— Content search and retrieval, domain ontology, ranking system, concept-focused crawling, eLearning.

I. INTRODUCTION

The pace of growth of the world-wide body of available information in digital format (text and audiovisual) constitute a permanent challenge for content retrieval technologies [1]. The popularity of exchange and dissemination of content through the web has created a huge amount of educational resources and the challenge of locating suitable learning references specific to a learning topic has become a big challenge [2]. As the web grows it will become increasingly difficult for educators to discover and aggregate collections of relevant and useful educational content. There is, as yet, no centralized method of discovering, aggregating and utilizing educational content [3].

This work proposes that Information Retrieval (IR) techniques and technologies could be specifically designed to traverse the WWW and centrally collect educational resources, categorized by topic area. IR systems are generally concerned with receiving a user's information need in textual form and finding relevant documents which satisfy that need from a specific collection of documents [3]. Most existing content retrieval techniques rely on indexing keywords. Unfortunately, keywords or index terms alone cannot adequately capture the document contents, resulting in poor retrieval performance [7]. Typically, the information need is expressed as a combination of keywords and a set of constraints. However, here we use learning terms associated with topic under consideration extracted from the domain ontology. These topics and learning terms are used in the concept based query method.

In addition, this work proposes a concept and term based ranking system for ordering the documents from

search engine to obtain a ranked list of seed documents. With the appearance of sophisticated search engines, finding materials for e- learning is not a problem. However, the resources that one discovers might have varying styles and may be targeted at different type of audiences. The resources may not have a complete coverage of topics which the instructor actually requires for content authoring. Moreover, a number of resources which are retrieved are highly redundant [4]. Hence, appropriate ranking of documents using concept and topic learning terms possibly will help in retrieving topic related documents and reducing redundancy from retrieved content. In this work, the ranking system exploits the concept-document similarity of the document collection. These ranked documents could then be used as seed documents for our proposed crawling system.

A focused crawler can be defined as a web crawler which actively seeks, acquires indexes and maintains pages on a specific topic which represent a relatively narrow segment of the WWW [14]. The goal of a concept-focused web crawler is to selectively seek out pages that are relevant to a predefined set of concepts associated with a topic. Besides sourcing web pages based on its content, concept-focused crawling allows a web crawler to process specific sites to greater depths than general purpose crawlers. Focused crawlers can spend more time perusing highly relevant sites rather than attempting to attain broad coverage of the entire WWW in a breadth-first manner [5]. During crawling, the crawling system described in this paper matches the ontology concepts with results. After crawling, concept-terms similarity ranking system has been used to rank the crawled documents. As a result, highly relevant pages can be discovered that may have been overlooked by

normal focused crawlers, at the same time filtering redundant pages to avoid unnecessary paths.

This paper is organized as follows: section II presents the proposed system for retrieval of educational content. Section III presents Ontology based query method for generating queries about concepts. Section IV presents Concept and term based ranking of documents. Section V presents Concept-focused crawling system.

II. PROPOSED SYSTEM

The figure.1 shows that the proposed system for retrieval of educational contents from web for e-learning consists of 3 phases. The first phase is the Ontology based query method for obtaining seed documents where we have been generated queries based on ontology concepts and used TF-IDF for deciding terms for query expansion. The inputs of this phase are ontology concepts which are given as queries to search engine. The output of this phase is set of documents which would be used as seed documents for crawling system. The second phase is the Concept-term based ranking of seed and crawled documents. The second phase describes the concept-term based ranking of documents. This phase takes a set of documents and ranks them according to concepts and terms associated with these concepts. The third and final phase is the Concept-focused crawling system which takes a set of ranked seed documents and produces a set of crawled documents. These documents are again ranked by using concept-term based ranking which gives final set of crawled documents.

Figure.1 Retrieval of educational content from web

III. ONTOLOGY BASED QUERY METHOD

While dealing with content retrieval from the web for educational contents, using the TF-IDF for determining terms for query expansion is inadequate and different methods are required to find out related concepts. In addition, a simple TF-IDF based re-querying scheme can leave out some potentially relevant sub concepts related with the query term [6]. An approach to the finding of educational content from the web is determined by a set of keywords in a file [3]. However, such an approach which is completely TF-IDF based would not be able to identify related words.

Similar to the work described by [11], the proposed system also used the concepts of the ontology to query the web to obtain seed documents. The ontology used by us is however specially designed a compute science ontology based on the ACM classification hierarchy. The association of terms to concepts for specific purposes has been used by InfoWeb [13] a filtering system using user profiles in a digital library scenario. Here the semantic network used to represent the user profile has nodes representing concepts and as more information is gathered about the user the profile is enhanced by associating additional weighted keywords with these concept nodes. This idea has been used in the work described in this paper where in the ontology, each node in addition to having concepts from ACM classification, has an associated set of topic learning terms typically used when teaching this topic. At present this set of associated topic learning terms is manually obtained from typical texts covering the topic. As a future enhancement we propose to enhance this ontology through machine learning techniques. The search using concepts and topic learning terms from the ontology retrieves a set of seed documents.

The ontology based query method scheme is required because it integrates some additional information about relationships. For example, when we give a query to the search engine for learning objects relevant to 'OOPS class' the keyword based search [3] can not identify "OOPS inheritance" as a related word because it is oblivious to the fact that a part-of relationship exists between an inheritance and a OOPS class. This ontology based query method involves a TF-IDF based querying to the search engines and also an ontology based query expansion for good content exposure from the web.

IV. CONCEPT-TOPIC LEARNING TERMS BASED RANKING

The majority of the resources retrieved from the web may be redundant. Presenting this kind of redundant materials to the learner may result in reduction in learner interest and efficiency. Identifying the most relevant results to a query is a central problem

in web search; hence ranking of documents has received a lot of attention. Mehrnoush Shamsfard et al, Azadeh Nematzadeh et al, and Sarah Motiee et al. presented an Ontology based system for ranking documents that combines conceptual, statistical and linguistic features of documents, and also expanded the query with its related concepts before comparing the documents [8]. However our concept-term ranking system which is oriented towards the retrieval of e-learning content and is based on two basic principles.

- i. *Documents having concepts of the topic along with learning terms, and sub concepts associated with these concepts are better resources for e-learning.*
- ii. *Documents which are about sub concepts of terms and not covered in the documents hitherto ranked are ranked as better resources - since we want to avoid redundancy of materials and also ensure better coverage of the topics.*

The ranking mechanism does four comparisons based on concepts and topic learning terms to rank documents. We explain the ranking mechanism using seed documents. This mechanism is also used to rank crawled documents. The number of concepts from the ontology that the seed document has, the similarity of concepts with concepts in other seed documents, number of topic learning terms associated with concepts and coverage of topic learning terms of the concept's sub-concepts all contribute to the ranking mechanism.

After obtaining the seed documents, we perform the ranking mechanism to select and crawl resources that are specifically suited to e-learning. However this type of ranking ensures coverage of topic concept, sub concepts and topic learning terms but does in no way evaluates the documents so retrieved from the view point of pedagogy or structure of learning content. The ranking mechanism has been used to rank the seed documents from search engine and also rank the crawled documents for filtering out appropriate documents for e-learning.

V. CONCEPT-FOCUSED CRAWLING SYSTEM

Focused crawler for intelligent web based content management system mined the educational contents from university websites in the form of course pages [9]. This system has mined only from university websites using Hidden Markov Models (HMMs) and Conditional Random Fields (CRFs). Jun Li et al, Kazutaka Furuse et al, Kazunori Yamaguchi et al proposed a method using a decision tree on anchor texts of hyperlinks for focused crawling. They have used feature selection for picking keywords for decision trees [10]. Marc Ehrig et al and Alexander Maedche et al proposed the focused crawling process consisting of two interconnected cycles: First, the ontology cycle, and second, the crawling cycle. In their work, the framework included a definition of the

knowledge structure and its lexical presentment, relevance computation strategies, an implementation, and an empirical evaluation of the overall framework [12].

Hiep Phuc Luong et al, Susan Gauch et al, Qiang Wang et al presented an ontology learning process framework that includes crawling, classifying and extracting relevant information from documents. They have used SVM classification to filter out documents in the searched results [11]. However our crawling system use the concept-term ranking to filter out the searched results.

Most of the above focused crawlers have used keywords alone for focused crawling system. However, our crawling system initially ranks seed documents so as to be able to crawl and obtain documents suitable for e-learning. We rely on background knowledge with concepts and associated topic learning terms, which are compared with the contents of the crawled documents. We use the concept-term ranking again to obtain the final set of crawled documents which is needed to avoid redundancy of content.

VI. CONCLUSION

Generic search engines and crawlers are like public libraries: they try to cater to everyone, and do not specialize in specific areas. In this paper we have presented concept-term based ranking system for obtaining seed documents from search engine and presented a concept-focused crawling system for the discovery of educational content from the web. The concept-term ranking and the concept-focused crawling system both have extracted the concepts and the topic learning terms from the domain ontology. We have proposed an algorithm for ranking the results obtained from web search based on the redundancy of content, so as to improve the variety of the content retrieved. This ranking system has been applied while searching for seed documents and also for concept-focused crawling system.

REFERENCES

- [1] David Vallet, Pablo Castells, Miriam Fernández, Phivos Mylonas, and Yannis Avrithis, "Personalized Content Retrieval in Context Using Ontological Knowledge", *Circuits and Systems for Video Technology*, IEEE Transactions on In Circuits and Systems for Video Technology, IEEE Transactions on, Vol. 17, No. 3. (05 March 2007), pp. 336-346.
- [2] Khairil Imran Bin Ghauth, Nor Aniza Abdullah, "Building an E-Learning Recommender System using Vector Space Model and Good Learners Average Rating", *Advanced Learning*

- Technologies, 2009. ICALT 2009 Ninth IEEE International Conference (15-17 July 2009), pp 194 - 196
- [3] Lawless, S. "Leveraging Content from Open Corpus Sources for Technology Enhanced Learning", Ph.D Thesis, Submitted to the University of Dublin, Trinity College, 2009.
- [4] Brusilovsky, P. & Henze, N. "Open Corpus Adaptive Educational Hypermedia". In *The Adaptive Web: Methods and Strategies of Web Personalisation*, Lecture Notes in Computer Science, vol. 4321, Berlin: Springer Verlag, pp. 671-696. 2007.
- [5] Chakrabarti, S., Punera, K., Subramanyam, M. "Accelerated Focused Crawling through Online Relevance Feedback". In *proceedings of the Eleventh International World Wide Web Conference, WWW2002*, Honolulu, Hawaii, USA. May 7-11, 2002.
- [6] Sparck-Jones, K (1972). "A statistical interpretation of term specificity and its application in retrieval", *Journal of Documentation* 2004, Volume 60 Number 5 pp. 493-502
- [7] DIK L. LEE, "Document Ranking and the Vector-Space Model", *Software, IEEE* (Mar/Apr 1997) Volume: 14 Issue: 2 pp 67 – 75.
- [8] Mehrnoush Shamsfard, Azadeh Nematzadeh, and Sarah Motiee, "ORank: An Ontology Based System for Ranking Documents", *International Journal of Computer Science*, Vol. 1, No. 3. (2006), pp. 225-231.
- [9] Udit Sajjanhar, "Focused Web Crawling for E-Learning Content", M.Tech Thesis to be submitted Indian Institute of Technology Kharagpur, April 2008
- [10] Jun Li Kazutaka Furuse Kazunori Yamaguchi, "Focused Crawling by Exploiting Anchor Text Using Decision Tree", *Proceeding WWW '05 Special interest tracks and posters of the 14th international conference on World Wide Web ACM New York, NY, USA 2005*, pp.1190-1191
- [11] Hiep Phuc Luong Susan Gauch Qiang Wang, "Ontology-based Focused crawling", *International Conference on Information, Process, and Knowledge Management*, Cancun, Mexico, Feb. 1-7, 2009, pp123-128.
- [12] Marc Ehrig Alexander Maedche, "Ontology-Focused Crawling of Web Documents", *Proceeding SAC '03 Proceedings of the 2003 ACM symposium on Applied computing*, ACM New York 2003
- [13] Gentili, G., Micarelli, A., Sciarone, F.: Infoweb: "An Adaptive Information Filtering System for the Cultural Heritage Domain". *Applied Artificial Intelligence* 17(8-9) (2003) 715-744
- [14] Chakrabarti, S., van den Berg, M., Dom, B. "Focused Crawling: A New Approach to Topic-Specific Web Resource Discovery". In *The International Journal of Computer and Telecommunications Networking*, Vol.31(11-16), Elsevier North-Holland, Inc. New York, USA. pp. 1623-1640, May 1999.

□□□