International Journal of Advances in Electronics Engineering – IJAEE
Volume 4 : Issue 1 *[ISSN 2278 – 215X]*
Publication Date : 09 January 2014

A WEB-AS-CORPUS APPROACH TO POPULATING WIKIVERSITY FOR TEACHING INFORMATION TECHNOLOGY MODULES

Bayan AbuShawar; Eric Atwell; Magdi Sawalha
IT department; School of Computing
Arab Open University; University of Leeds
Amman, Jordan; leeds, Uk
b_shawar@aou.edu.jo; eric@comp.leeds.ac.uk;scmss@leeds.ac.uk

***Abstract*—Wikiversity is an online open-source public repository for University-level teaching and learning materials, based on the Wikipedia architecture for "crowd-sourcing": it relies on volunteers to collaborate by actively contributing their knowledge for the common good. Undoubtedly a huge volume of learning resources exists on the WWW, but scattered on individual websites, in a wide range of formats and structures. Individual lecturers prepare online teaching materials to support their own teaching, but few know of Wikiversity, and few have time or inclination to take on the extra step of formally registering and uploading their materials to Wikiversity. In this paper we will present an approach to organise and semi-automate the harvesting of these scattered resources, by using BootCat to crawl the Web and adapting Web-as-Corpus techniques. Results show that scouting the Web returns better resources than BootCat toolkit which returns a lot of irrelevant links such as forums, and personal repositories.**

***Keywords*- wikiVersity; BootCat; Web-as-Corpus.**

I. INTRODUCTION

Wikipedia is "a free encyclopedia, written collaboratively by the people who use it. It is a special type of website designed to make collaboration easy, called a wiki." Wiki is used as a teaching tool; it enables learners to create, edit, and communicate, and synthesized subject knowledge in a shared and open space. [1].

Wikipedia is a free, collaborative, multilingual Internet encyclopedia supported by the non-profit Wikimedia Foundation. Wikipedia was launched in January 2001 [2]. Wikipedia's departure from the expert-driven style of encyclopedia building and the presence of a large body of un-academic content has received ample attention in print media. In its 2006 Person of the Year article, Time magazine recognized the rapid growth of online collaboration and interaction by millions of people around the world [3]. Wikipedia has also been praised as a news source because of how quickly articles about recent events appear [4-5].

Wikiversity *is one of the sister projects of Wikipedia. Wikiversity is an online open-source public repository for Universal-level teaching and learning materials, based on the Wikipedia architecture for "crowd-sourcing". Wikiversity has mainly three aims [6-7]:*

- To create, and host different educational materials for all groups in different languages;
- To develop learning activities and communities: learning community, service community, and a research community.
- To facilitate research projects and host research results.

Individual lecturers prepare online teaching materials to support their own teaching, but few know of Wikiversity, and few have time to take on the extra step of formally registering and uploading their materials to Wikiversity. In this paper we present an approach to organise and semi-automate the harvesting of these scattered resources, by adopting Web-as-corpus techniques and using BootCat to crawl the Web. Background is described in section 2, Section 3 presents BootCat toolkit, the policy of Arab Open University is and its needs for extra resources is shown in section 4, section 5 discusses the harvesting approaches for IT materials, evaluations and conclusions are presented in

SEEK DIGITAL LIBRARY

section 6 and 7 consequently.

II. WEB AS A CORPUS

Language and Linguistics research and teaching often involves a Corpus or empirical sample of the language being studied.

Many areas of linguistics benefit from empirical corpus evidence, including syntax, semantics, pragmatics, and discourse. There are no constraints on the size, or contents of a corpus. A corpus could contain the entire work of an author, or all questions and answers related to specific domain, or all historical information about a piece of art. A corpus could be extracted from written text as well as transcribed from recorded speech.

The WWW is increasingly used as a corpus source, at three levels:

1- A web search engine such as Google or Yahoo allows linguists to view the WWW as a corpus, by searching for and extracting empirical evidence of language items directly from the Web.
2- A "robot" can be used to automate the process of harvesting a corpus fitting specified criteria; for example BootCat[8] is given a list of specific "seed terms" and uses these to harvest WWW documents matching some or all of these seed terms.
3- Scholars can "scout" the Web for suitable text-samples fitting their research needs, and download them to compile a specialised corpus.

We propose to adapt and combine these three complementary approaches, to harvest Wikipedia-level teaching and learning resources for teaching and learning about Computing Studies.

As a sample of our experiment, some of IT modules at Arab Open University (AOU), and School of Computing at University of Leeds will be used to extract significant terminology that will be used during searching process to enrich Wikiversity site.

III. BOOTCAT TOOLKIT

Several researchers used bootstrapping techniques in building their corpora and extracting terms such as BootCat[8] and CrawlTCQ[9].

BootCat[8] is a "suit of Perl programs implementing an iterative procedure to bootstrap specialised corpora and terms from the Web, requiring only a small list of seeds as input".

Fantinuoli [10] uses BootCat to build a corpus then compares between manual and automatically constructed corpora by comparing results of a terminological extraction from the two corpora. Results prove the usefulness of using BootCat and the benefit from reducing time an effort in manual process. Dillon [11] uses BootCat to gather corpora of academic writing from the Web. Ghani et al. [12] has built minority language corpora from the Web. Sharoff [13], Ferraresi et al. [14], and Baroni et al. [15] built very large web-derived corpus in various languages based on WaCky Wide Web.

The basic procedure of BootCat toolkit is as follow [8]:

1- The Bootstrapping process starts with a small list of terms that are significant to domain under investigation. The seed terms can be either single words, or multi terms enclosed in double-quotes.
2- The seed terms are randomly combined and each combination is used as Google query string.
3- The top n pages returned for each query are retrieved and formatted as text.

This is the first round in building the specialist corpus, new unigrams seeds are extracted from the corpus of retrieved pages, random combinations of the newly extracted seed terms are then used for another round of Google quires and a new corpus is created, this process is repeated to enlarge specialised corpus as needed. There are a lot of parameters that are determined by the user:

- The number of iterations;
- The number of queries issued for each iteration;
- The number of seeds used and number of pages to be retrieved are determined by the user.

A new Web-server version of BootCat (WebBootCat) [16] is widely used nowadays, where there is no need to download or install software, and which is easier to use for non-technical people.

IV. HARVESTING TEACHING MATERIAL FOR IT: APPROACHES, RESULTS, AND EVALUATION

We propose to harvest Wikipedia-level

SEEK DIGITAL LIBRARY

teaching and learning resources for teaching and learning about Computing Studies. We investigated and selected some of Information Technology modules and terms to identify candidate teaching materials, this was done by some of the staff who are familiar with those modules, the syllabus of each module was examined for significant lexical.

Using the "search engines" to surf the internet for suitable teaching material is time consuming, so the extracted lexical terms were used to adapt the Web-as-corpus paradigm at two levels:

1- Using these lexical terms, to "Scout" Google, view and extract required material.
2- Adapting BootCat, the web-as-corpus robot, to automatically harvest potential information technology teaching materials using the same extracted lexical terms.

A. *Approach One: Scouting the Web*

A sample module was selected from IT department at AOU named: an object oriented programming with Java, M255. This module consists of 14 units, where each unit starts by listing the most important topics aimed to be covered in the unit. We use the list of each unit to extract the most significant terms and phrases as shown in table 1.

Approach one was applied, scouting Google for PowerPoint slides related to Java programming. We believe that most of lecturers prefer using PowerPoint slides in their teaching, so we hope to gain more material which presenting Java topics in simple way. The general Google query is: * Java .ppt

For example: inheritance in Java .ppt

The first 10 top hit pages were retrieved, in total 55 files were saved after eliminating duplications. An expert, who is teaching the module, evaluates the relevance of these files to M255 module. Five of these files were deleted since C++ language is used to present object oriented concepts.

TABLE 1. M255 LEXICAL TERMS

	Java Lexical terms and phrases
1	Object oriented concepts
2	Access modifiers
3	Abstract class and interface
4	Collections
5	Sets
6	Maps
7	Lists
8	Arrays
9	Polymorphism
10	Static variables and methods
11	I/O streams
12	Reading/writing files
13	Inheritance
14	Exceptions and errors
15	Control structure

In another round, scouting the web approach was tested using some lexical and phrases extracted from AI123 module, which is an Artificial Intelligent module that is taught in School of Computing at University of Leeds, such as: algorithms for semantic analysis; word sense disambiguation; corpus-based approaches; machine-learning approaches; tagging; data-driven approaches.

Most of related top hit pages were Wikipedia ones which are easy to understand by non-expert. In this round, the selected seed terms were used to look up in Wikipedia to yield tutorial-style textbook material.

We scout the Web for Wikipedia pages, extract the suitable text from Wikipedia and create Wikipedia textbook for A123 module.

B. *Approach Two: Using BootCat*

BootCaT [8] toolkit implement an iterative procedure to bootstrap specialized corpora and terms from the web, requiring only a list of "seeds" (i.e. terms that are expected to be typical of the domain of interest) as input. In this manner, the same lexical used in table 1 for M255 module were tried as seed terms in BootCat software, however, as a first round, no tuples were generated in order to search for the same seeds,

SEEK DIGITAL LIBRARY

and no restriction on domains, as a result 93 links were retrieved as shown in figure 2.

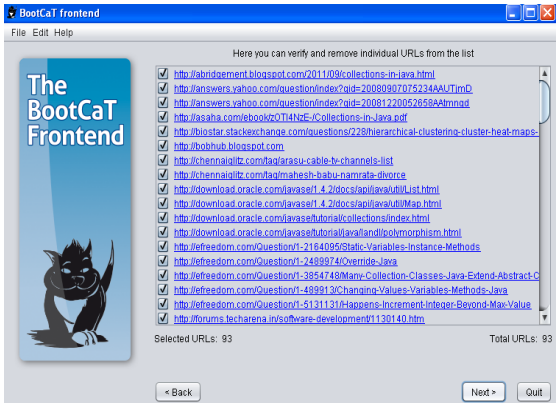

Figure 2. Retrieved URLS of M255 module from BootCat

An initial evaluation applied by an expert before constructing the corpus, he URLS were checked and the following links were eliminated:

- Forum web pages
- Personalise repositories
- WebPages that answers users questions (http://answers.yahoo.com/question)
- Some WebPages were blocked
- Irrelevant WebPages such containing C++ Object oriented.

As a result 74 links were irrelevant links, and 19 of WebPages were selected to generate M255 corpus. BootCat was used to crawl the Web for AI123, however the generated corpus contained a lot of irrelevant resources.

In parallel to these approaches to "harvesting", an evaluation procedure is developed, to filter out unsuitable candidate teaching resources. Potential resources are reviewed by academic teaching staff in IT department, who will use their expertise and judgement to evaluate (i) relevance to their own teaching, and (ii) potential usefulness to the wider community of teachers and learners computing.

We are teaching the modules that were selected, an initial evaluation for scouting the Web shows that the retrieved resources are relevant.

However in case of using BootCat, a lot of retrieved links were irrelevant, in this case, changing some of the parameters, like number of tuples, and restrict searching to specific domains is required to make resulted corpus more appropriate. In a second round of using BootCat with AI123 seed terms, we tried to build a corpus without using any tuples and with limiting domain to .org, in order to retrieve Wikipedia pages. The generated corpora were much better than the previous one.

V. Conclusion and Future Work

Wikiversity is suitable for publishing the collected material to be widely used by learner and teacher community. In our methodology, we cited the course syllabus, as a source of keywords for searching the Web, via BootCat and then via "scouts" in order to enrich Wikiversity by extra educational materials.

Initial evaluations show that scouting the Web returns better results than using BootCat. Most of the returned links from BootCat were Wikipedia pages; there is no way to restrict searching for power point slides lectures for example. Although the aim of this experiment is not to compare webCat with scouting, but we found that Wikipedia could be used to source material for a "Wiki-textbook": a customised collection of teaching materials suited to the specific topics of a course.

In future work we can use the keywords to find matching articles in Wikipedia, and this allowed us to harvest relevant Wikipedia pages written by experts in a tutorial style to be read by non-experts; ideal for students!. We can also see how to apply this method to produce personalised wiki-textbook teaching materials for specific university-level syllabuses.

VI. References

[1] M.-F.G Lin; S Sajjapanroj; and C.J Bonk. "Wikibooks and Wikibookians: Loosely Coupled Community or a Choice for Future Textbooks?" IEEE Transactions on Learning Technologies, 4(4), 327-339, 2011.

[2] M. Miliard. Wikipediots: Who Are These Devoted, Even Obsessive Contributors to Wikipedia? Salt Lake City Weekly. 2008.

[3] G. Lev. Time's Person of the Year: You. TIME (Time, Inc), 2006.

[4] L Andrew. Wikipedia as Participatory Journalism: Reliable Sources? Metrics for Evaluating Collaborative Media as a News Resource. 5th International Symposium on Online Journalism (University of Texas at Austin), 2004.

[5] D. Jonathan. All the News That's Fit to Print Out. The New York Times Magazine, 2007.

[6] N. Friesen, J. Hopkins. "Wikiversity; or education meets the free culture movement: An ethnographic investigation". First Monday, 13(10), 1-14, 2008.

[7] C. Lawler. "Action Research as a Congruent Methodology for Understanding Wikis: The Case of Wikiversity". JIME http://jime.open.ac.uk, pp. 1-11.

SEEK DIGITAL LIBRARY

[8] M. Baroni; S. Bernardini. BootCat: Bootstrapping Corpora and Terms from the Web. In Proceedings of the LREC 2004 conference.

[9] C. Groc; A. Tannier; J. Couto . GrawlTCQ: Terminology and Corpora Building by Ranking Simultaneously Terms, Queries and Documents using Graph Random Walks. In Proceedings of the TextGraphs-workshop, 37-41, 2011.

[10] C. Fantinuoli. "Specialized corpora from the Web and term extraction for simultaneous interpreters". M. Baroni, S. Bernardini (eds.) WaCky! Working papers on the Web as Corpus, Bologna, 173-190, 2006.

[11] G. Dillon G. Building Webcorpora of academic prose with BootCat. In Proceedings of the NAACL HLT 2010 Sixth Web as Corpus Workshop, 2010, 26-31.

[12] R. Ghani; R. Jones; D. Mladenic. "Building Minority Language Corpora by Learning to Generate Web Search Queries. Knowl". Inf. Syst., 7(1),56–83, 2005.

[13] S. Sharoff. "Creating general-purpose corpora using automated search engine queries". M. Baroni, S. Bernardini (eds.) WaCky! Working papers on the Web as Corpus, Bologna, 63–98, 2006.

[14] A. Ferraresi; E. Zanchetta; M. Baroni; S. Bernardini. Introducing and evaluating ukwac, a very large web-derived corpus of English. In Proceedings of the 4th Web as Corpus Workshop (WAC-4), 47–54, 2008.

[15] M. Baroni; S. Bernardini; A. Ferraresi; A. Zanchetta. TheWaCkyWideWeb: A Collectionof Very Large Linguistically Processed Web-Crawled Corpora. In Proceedings of the LREC 2009 conference, volume 43, 209–226, 2009.

[16] M. Baroni; A. Kilgarriff; J. Pomikalek; P. Rychly WebBootCat: a web tool for instant corpora. In Proceeding of the EuraLex Conference 2006, 123-132, 2006.

SEEK DIGITAL LIBRARY