OERScout: Autonomous Clustering of Open Educational Resources using Keyword-Document Matrix

Ishan Sudeera Abeywardena[a], Choy Yoong Tham[b], Chee Seng Chan[c] and Venkataraman Balaji[d]

[ab] School of Science and Technology, Wawasan Open University, 54 Jalan Sultan Ahmad Shah, Penang, 10050, Malaysia. e-mail: [a] ishansa@wou.edu.my , [b] cytham@wou.edu.my

[c] Faculty of Computer Science and Information Technology, University of Malaya, 50603, Kuala Lumpur, Malaysia. e-mail: cs.chan@um.edu.my

[d] Commonwealth of Learning (COL), 1055 West Hastings Street, Suite 1200, Vancouver, BC V6E 2E9, Canada. e-mail: vbalaji@col.org

Sub-theme:
Open Educational Resources (OER) and ODL

Abstract

The Open Educational Resources (OER) movement has gained momentum in the past few years. With this new drive towards making knowledge open and accessible, a large number of OER repositories have been established and made available online throughout the globe. However, despite the fact that these repositories hold a large number of high quality material, the use and re-use of OER has not taken off as anticipated due to various geographic, socio and technological limitations. One such technological limitation is the present day inability to effectively search and locate OER materials which are specific and relevant to a particular academic domain. As a first step towards a possible solution to this issue, this research paper discusses the design and development of a clustering algorithm which accurately clusters text based OER materials by building a Keyword-Document Matrix (KDM) using autonomously identified domain specific keywords. This algorithm is the first phase of a larger technology framework named "OERScout" which is a potentially new methodology for effectively searching and locating desirable OER for academic use.

Keywords: OERScout, Open Educational Resources, OER, OER searching and location, Text mining algorithms, Document clustering, Autonomous keyword identification

1 Introduction

With the new drive towards accessible and open information, Open Educational Resources (OER) have taken centre stage after being first adopted in a UNESCO forum in 2002. OER can be defined as "*web-based materials, offered freely and openly for use and re-use in teaching, learning and research*" (Joyce, 2007) which are heavily dependent on technology and the internet to be accessible by the masses. According to Farber (2009) "*Just as the Linux operating system and other open source software has become a pervasive computer technology around the world, so too might OER materials become the basis for training the global masses*" which clearly outlines the significance of OER as a global movement. The move towards OER has also helped reduce significantly the costs of production, reproduction and distribution of course material (Caswell, Henson, Jenson & Wiley, 2008) especially as initiatives such as MIT OpenCourseWare (OCW), Rice University Connexions and the Commonwealth of Learning (COL) funded Wikieducator project are sharing high quality educational resources under the Creative Commons (CC) license which enables institutions and individuals globally to adapt and re-use material without developing them from scratch. This is especially important for countries in the Global South such as India which has 411 million potential students, out of which only 234 million enter school at all, less than 20% reach high school and less than 10% graduate (Kumar, 2009).

Over the recent past, many global OER initiatives have been established by organisations such as UNESCO, COL and the United Nations (UN) to name a few. Many of these initiatives are based on established web based technology platforms and have accumulated large volumes of quality resources which are shared with the masses. However, the use of diverse and disparate technology platforms in these projects entails the inability to effectively trawl and located OER using generic search methodologies. This is affirmed by Abeywardena, Raviraja and Tham (2012) who state that there is still no generic methodology available at present to enable search mechanisms to autonomously gauge the *desirability* of an OER which is a function of (i) the level of openness; (ii) the level of access; and (iii) the relevance; of an OER for ones needs. Thus, the necessity for a methodology which could effectively trawl and search the numerous disconnected and disparate OER repositories with the aim of locating *desirable* materials has taken center stage as the problems with open content is not the lack of available resources on the Internet but the inability to locate suitable resources for academic use (Unwin, 2005).

OERScout is a technology framework which aims to accurately cluster text based OER materials by building a Keyword-Document Matrix (KDM) using autonomously mined domain specific keywords. Using the KDM, the system accurately generates lists of specific and relevant OER from the distributed repositories to suit a given search query. In this context, *specific* denotes the suitability of an OER for a particular teaching need. For example, an OER on physics from the final year syllabus of a physics degree would not be suitable for a high school physics class; and *relevant* denotes the match between the content of the OER and the content needed for a particular teaching need. For example, physical chemistry is not relevant for a teaching need in organic chemistry. This paper, which is organised under the headings *methodology*, *pilot tests*, *discussion* and *conclusion*; discusses how *OERScout* benefits the ODL community, who are arguably the largest group of OER creators and consumers (Abeywardena, 2012), by providing a centralised system for effectively searching and locating specific and relevant OER materials form the disconnected and disparate repositories scattered across the globe.

2 Methodology

The *OERScout* text mining algorithm was designed to "read" text based OER and "learn" which academic domain(s) and sub-domain(s) they belonged to. To achieve this, a *bag-of-words* approach was used due to its effectiveness when used with unstructured data (Feldman & Sanger, 2006). The algorithm extracted all the individual words from a particular text by removing noise such as formatting tags and punctuations to form the corpus. The corpus was then *Tokenised* into the *List of Terms* using the *stop words* found in the Onix Text Retrieval Toolkit[1]. The extraction of the meaningful terms from the *List of Terms* for the formation of the Term Document Matrix (TDM) was done using the Term Frequency (TF) – Inverse Document Frequency (IDF) weighting scheme. The weight of each term (TF-IDF) was calculated using the following formula where *N* is the number of documents in the collection; TF_t is the frequency of a term *t* in a single document; and IDF_t is the frequency of a term *t* in all the documents in the collection [IDF_t= Log (N/TF_t)].

$$(TF\text{-}IDF)_t = TF_t \text{ x } IDF_t$$

The probability of a term *t* being able to accurately describe the content of a particular OER as a keyword decreases with the number of times it occurs in other related and non-related materials. For example the term "introduction" would be found in many OER which discuss a variety of subject matter. As such the TF-IDF of the term "introduction" would be low compared to a term such as "operating systems" or "statistical methods" which are more likely to be keywords. As the TF-IDF weighting scheme takes the inverse document frequency into consideration, it was found to be suitable for extracting the keywords from an OER.

The formation of the Keyword-Document Matrix (KDM) was done by (i) normalising the TF-IDF values for the terms in the TDM; and (ii) applying the Pareto principle (80:20) where the top 20% of the TF-IDF values were considered to be keywords describing 80% of the OER (Figure 1).

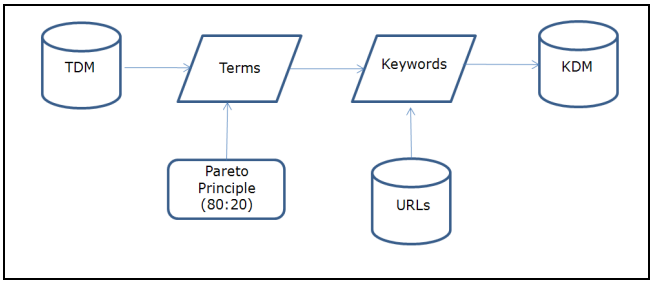

Figure 1 Creation of the KDM

The *OERScout* algorithm was implemented using the Microsoft Visual Basic.NET 2010 (VB.NET 2010) programming language. The corpus, *List of Terms*, TDM and KDM were implemented using the Microsoft SQLServer 2008 database platform. The OER resources were fed into the system using sitemaps based on extensible markup language (xml) which contained the uniform resource locators (URLs) of the resources.

[1] lextek.com/manuals/onix/stopwords1.html

3 Pilot Tests

Two pilot tests were conducted to test the functionality of the system. As the first test case, the Rice University's OER repository Connexions[2] was used due to (i) the large number of diverse OER materials available; (ii) the relatively high popularity and usage rates; and (iii) the availability of the OER materials in text format. An xml sitemap containing 1238 URLs belonging to the domains of arts, business, humanities, mathematics and statistics; science and technology; and social sciences was created as the initial input. The system was run with the initial input and was allowed to autonomously create the KDM. The average time taken for *OERScout* to extract terms from an OER and update the KDM was found to be approximately two minutes. After the completion of the pilot test, the system had created 1013 clusters in the KDM with an average density of 1.23 resources per cluster. It was also noted that 1238 resources had contributed 141901 new terms. An example of the KDM is shown in Figure 2.

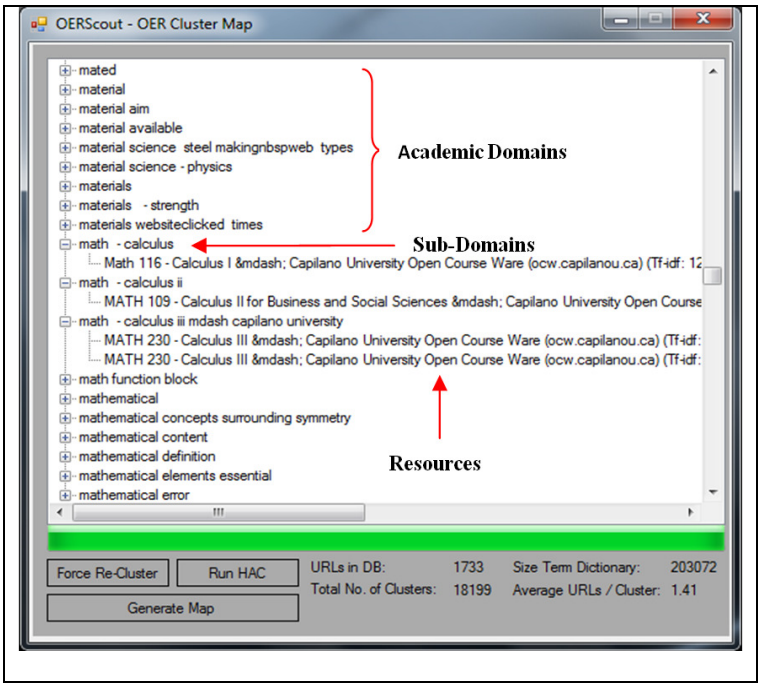

Figure 2 Example of the cluster map generated using the KDM

The second test was conducted on the Directory of Open Educational Resources (DOER)[3] of the COL. DOER is a fledgling *portal OER repository* (McGreal, 2010) which provides an easily navigable central catalogue of OER scattered across the globe. At present the OER available through DOER are manually classified into 20 main categories and 1158 sub-categories. However, despite covering most of the major subject categories, this particular ontology would need to expand by a large degree due to the unlimited variety of OER available in a kaleidoscope of subject areas. This expansion, in turn, becomes a tedious and laborious task which needs to be accomplished manually on an ongoing basis. As a possible solution to this issue, a mechanism was needed for autonomously identifying the subject area(s) covered in a particular OER, in the

[2] http://www.cnx.org
[3] http://doer.col.org/

form of keywords, in order for it to be accurately catalogued. Given this requirement DOER was used as the training dataset for the second pilot test of *OERScout*. This training process was critical to the functioning of the algorithm as it had to learn a large array of academic domains and sub-domains before being able to accurately cluster resources according to the domain. After completion of the second test, the system had processed 2598 resources of file types HTML, PDF, TEXT and MS Word from a multitude of OER repositories. On average, each resource required approximately 15-90 minutes to be read and learnt by the system. The creation of the KDM required approximately 12-24 hours each time.

3 Discussion

Generic search methodologies such as Google, Yahoo! and Bing are the most widely used search mechanisms for locating OER (Abeywardena & Dhanarajan, 2012). Even though this method is the most commonly used, it is not the most effective as discussed by Pirkkalainen and Pawlowski (2010) who argue that "*searching this way might be a long and painful process as most of the results are not usable for educational purposes*". Despite semantic web based alternatives such as Agrotags (Balaji et al., 2010) which build ontologies of domain specific keywords to be used for classification of OER belonging to a particular body of knowledge, the creation of such ontologies for all the domains discussed within the diverse collection of OER would be next to impossible. As such, the *OERScout* system was developed to use clustering techniques instead of semantic web techniques to enable OER to be clustered based on autonomously identified keywords.

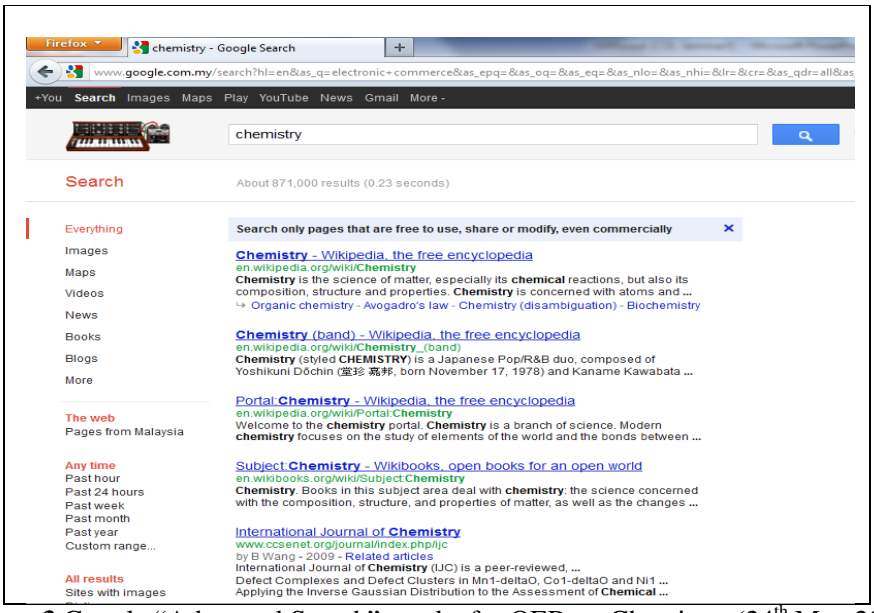

Figure 3 Google "Advanced Search" results for OER on Chemistry (24th May 2012)

Figure 3 shows an advanced search conducted on Google[4] for the term "chemistry" specifically searching for resources which are *free to use, share or modify, even commercially*. This example confirms the statements made in literature as the first three results are from Wikipedia[5] which is

[4] http://www.google.com.my
[5] http://www.wikipedia.org

an encyclopedia of user created learning objects rather than a repository of pedagogically sound educational material. Furthermore, the fifth result is a non-OER source. According to Vaughan (2004) users will only consider the top ten ranked results for a particular search as the most relevant. Vaughan further suggests that the users will ignore the results below the top 30 ranks. As such, generic search methodologies such as Google are currently inapt at locating specific and relevant OER for a particular teaching need.

Figure 4 shows a search result for the term "chemistry" on *OERScout* conducted on the KDM created during the second pilot test. Contrary to the static list of search results produced by typical search engines, *OERScout* provides an autonomously identified dynamic list of *Suggested Topics* which are related to "chemistry". The user is then able to click on any of the suggested topics to access specific and relevant OER, identified in the KDM, from all the repositories indexed by *OERScout*. Furthermore, based on the selection by the user, the system will provide a list of *Related Topics* which will enable the user to drill down further to identify the most suitable OER for his/her teaching needs. As such, it can be seen that *OERScout* is centralised system which is much more dynamic and effective in locating specific and relevant OER from the disconnected and disparate repositories. This becomes one of the major benefits to ODL practitioners as the system spares the user from conducting countless keyword searches in the OER repositories in order to identify suitable material for use. It also allows content creators to quickly isolate the OER suitable for their needs without reading through all the search results returned by a typical search mechanism such as Google.

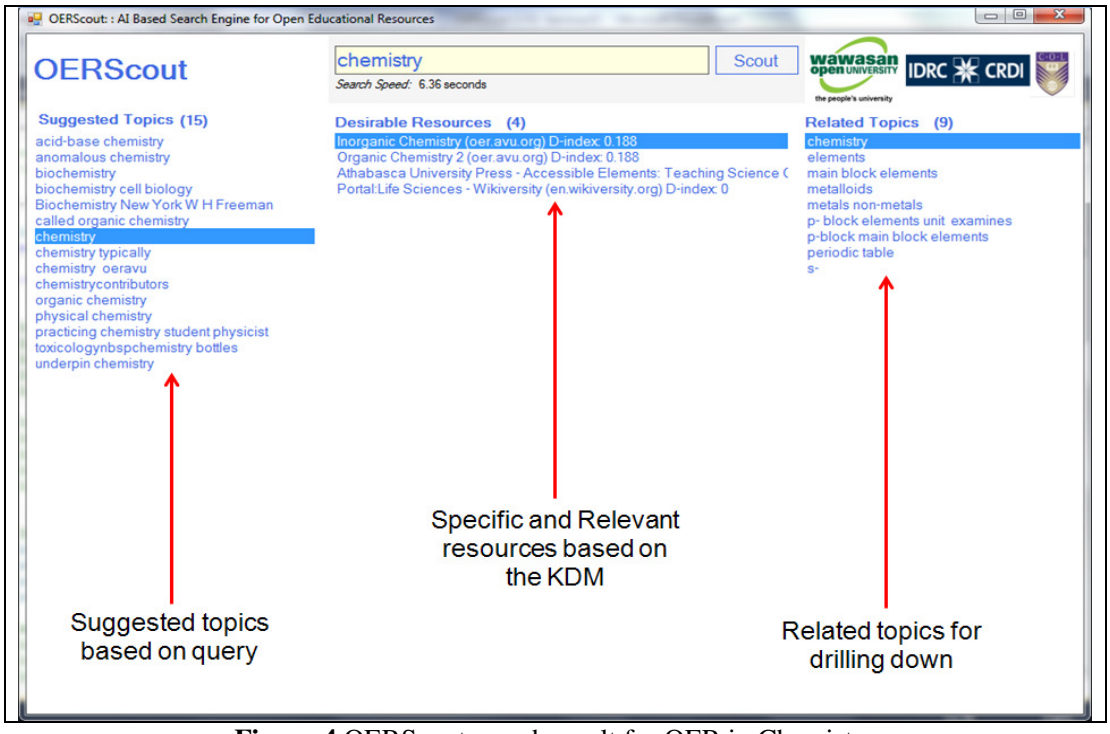

Figure 4 OERScout search result for OER in Chemistry

This first version of *OERScout* is unable to cluster non-text based materials such as audio, video and animations which is a major drawback considering the fact that more and more OER are now being developed in multimedia formats. However, it was noted from the pilot tests that the

system will accurately cluster multimedia based material using the text based descriptions provided. Another limitation is its inability to cluster resources written in languages other than English. Despite this current limitation, the *OERScout* algorithm has a level of abstraction which allows it to be customised to suit other languages in the future.

4 Conclusion

Open Educational Resources (OER) is a phenomenon which is rapidly gaining acceptance and credibility in the academic community as a potent tool for teaching and learning. With more and more OER repositories mushrooming across the globe and with the expansion of existing repositories due to increased contributions, the task of searching and locating specific and relevant OER has become a daunting one. This is further heightened due to the disconnectedness and disparity among the various OER repositories which are based on a number of technological platforms. Another hurdle to the searching and location of OER is the inability of current mainstream search technologies to effectively locate OER material for academic use. As such, each OER repository has to be searched using its own native search methodologies in order to locate the necessary OER. This again has had a discouraging effect on the OER practitioner as the number of repositories available is substantial.

OERScout is a text mining algorithm used for clustering OER using autonomously mined domain specific keywords. It was developed with a view of providing OER creators and users with a centralised system which will enable effective searching and location of specific and relevant OER for academic use. The benefits of *OERScout* to the content creators include (i) elimination of the need for manually defining content domains for categorisation in the form of metadata; (ii) elimination of the need for publicising the availability of a repository and the need for building custom search mechanisms for them; and (iii) more visibility and reach of material to a wider audience. The system benefits OER users by (i) providing a central location for finding resources scattered across the globe hidden in high volume repositories; and (ii) locating only the most specific and relevant resources. The ultimate benefit of *OERScout* is that both content creators and users now only need to concentrate on the actual content and not the searching and location of specific and relevant OER.

The next version of *OERScout* will enable ODL practitioners to effectively locate the most *desirable* OER for academic use based on parametric measures of (i) *openness* calculated using the Creative Commons license; (ii) *accessibility* calculated using the accessibility of the file format; and (iii) *relevance* calculated using the KDM.

Acknowledgements

This research project is funded as part of a doctoral research through the Grant (# 102791) generously made by the International Development Research Centre (IDRC) of Canada through an umbrella study on Openness and Quality in Asian Distance Education.

The development of the *OERScout* system was partially facilitated by the Commonwealth of Learning (COL), Canada through an Executive Secondment (4th – 25th May 2012).

This research paper is partially supported by Grant-in-Aid for Scientific Research (A) to Tsuneo Yamada at the Open University of Japan (JSPS, Grant No. 23240110).

Ishan Sudeera Abeywardena acknowledges the support provided by the Faculty of Computer Science and Information Technology, University of Malaya, 50603, Kuala Lumpur, Malaysia where he is currently pursuing his doctoral research in Computer Science and the School of Science and Technology, Wawasan Open University, 54 Jalan Sultan Ahmad Shah, 10050, Penang, Malaysia where he is currently employed.

References

Abeywardena, I. S. (2012). A report on the Re-use and Adaptation of Open Educational Resources (OER): An Exploration of Technologies Available. *Commonwealth of Learning,* 51. Retrieved August 11, 2012 from http://www.col.org/resources/publications/Pages/detail.aspx?PID=411.

Abeywardena, I. S., & Dhanarajan, G. (2012). OER in Asia Pacific: Trends and Issues. *Keynote address of the Policy Forum for Asia and the Pacific: Open Education Resources organised by UNESCO Bangkok and Commonwealth of Learning (COL), 23rd April 2012, Thailand.* Report available at http://www.unescobkk.org/education/ict/online-resources/databases/ict-in-education-database/item/article/oer-in-asia-trends-and-issues/.

Abeywardena, I.S., Raviraja, R., & Tham, C.Y. (2012). Conceptual Framework for Parametrically Measuring the Desirability of Open Educational Resources using D-index. *International Review of Research in Open and Distance Learning, 13(2),* 104-121.

Balaji, V., Bhatia, M. B., Kumar, R., Neelam, L. K., Panja, S., Prabhakar, T. V., Samaddar, R., Soogareddy, B., Sylvester, A. G., & Yadav, V. (2010). Agrotags – A Tagging Scheme for Agricultural Digital Objects. *Metadata and Semantic Research Communications in Computer and Information Science 108*, 36-45.

Caswell, T., Henson, S., Jenson, M., & Wiley, D. (2008). Open Educational Resources: Enabling universal education. *International Review of Research in Open and Distance Learning 9*(1), 1-11.

Farber, R. (2009). Probing OER's huge potential [Electronic Version]. *Scientific Computing 26*(1), 29-29.

Feldman, R., & Sanger, J. (2006). The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data. *Cambridge University Press*, Pages???.

Joyce, A. (2007). OECD Study of OER: Forum Report, *OECD*. Retrieved December 12, 2011 from http://www.unesco.org/iiep/virtualuniversity/forumsfiche.php?queryforumspages_id=33.

Kumar, M. S. V. (2009). Open educational resources in India's national development. *Open Learning: The Journal of Open and Distance Learning 24*(1), 77-84.

McGreal, R. (2010). Open Educational Resource Repositories: An Analysis. *Proceedings: The 3rd Annual Forum on e-Learning Excellence, 1-3 February 2010, Dubai, UAE, Retrieved December 27, 2011 from http://elexforum.hbmeu.ac.ae/Proceeding/PDF/Open%20Educational%20Resource.pdf.*

Pirkkalainen, H., Pawlowski, J. (2010). Open Educational Resources and Social Software in Global E-Learning Settings. *In Yliluoma, P. (Ed.) Sosiaalinen Verkko-oppiminen. IMDL, Naantali, 23–40*.

Unwin, T. (2005). Towards a Framework for the Use of ICT in Teacher Training in Africa. *Open Learning 20*, 113-130.

Vaughan, L. (2004). New measurements for search engine evaluation proposed and tested. *Information Processing and Management 40*, 677–691.

Wolfenden, F. (2008). The TESSA OER Experience: Building sustainable models of production and user implementation. *Journal of Interactive Media in Education*. Retrieved December 9, 2011 from http://jime.open.ac.uk/2008/03/.

Onix Text Retrieval Toolkit API Reference. Retrieved December 19, 2011 from http://www.lextek.com/manuals/onix/stopwords1.html.