AN EFFICIENT EDUCATIONAL MATERIAL EXPLORATION USING EXTRACTED CONCEPTS

Tetsuro Takahashi[1,2] and Richard C. Larson[2]
[1] *Fujitsu Laboratories LTD., 4-1-1 Kamikodanaka, Kawasaki 211-88, Japan*
[2] *Sociotechnical Systems Research Center, Massachusetts Institute of Technology, Cambridge, Massachusetts U.S.A.*

Keywords: Educational Data Mining, Open Educational Resources, Concept Network, Natural Language Processing.

Abstract: There are huge amount of Open Educational Resources (OER) provided by academic institutes now. Learners have an advantaged environment in which they can use the educational materials for free in their learning. General text search techniques, however, are not enough to provide an efficient search function for them. This drawback makes it difficult for learners to fully exploit the promising OER. We propose to solve the problem by providing an environment in which learners can search for educational materials efficiently using a set of concept networks derived from the existing OER. The result of experiment with the actual OER shows the availability of the proposed method.

1 INTRODUCTION

In the decade following the start of sensational educational project, Open Courseware (OCW) leaded by Massachusetts Institute of Technology, many colleges and universities ("institutes" here after) have started to provide their educational resources for free on the WWW (Carson, 2009). There are more than 15,000 courses offered in 20 languages through the Open Courseware consortium[1].

iTunes University is another major platform for Open Educational Resources (OER). It is an Internet service offered by Apple, Inc. that allows institutes to share media files among students and faculty, as well as with the general public (Reid, 2008). Currently more than 350,000 educational materials are provided by about 800 institutes.

In general, educational materials has the following hierarchy and each lecture has several materials.

Institute → Subject → Course → Lecture,

Though the educational materials are provided in several formats such as video, transcript, lecture note, slide, exam and so on, we call them all "educational material" in this paper. For instance, MIT OCW has a subject "Mathematics" which has about 130 courses such as "Single Variable Calculus" or "Linear Algebra". Each course has about 30 lectures in average, and a lecture has several educational materials.

[1]http://www.ocwconsortium.org

Learners have an advantaged environment in which they can use the educational materials for free in their learning. However, there are several difficulties to utilize the OER efficiently and effectively. We discuss the difficulties in the next section.

2 DIFFICULTIES IN THE USE OF OER

2.1 Cross-institute

Every institute has their own subjects and curricula that the provided OER are following. While the OER have rigidly structured within an institute, the structures are not universal to all institutes. For example, a course "Single Variable Calculus" in an institute may or may not comparable to a course "Calculus" in another institute. Moreover it is not obvious which subject should a course "Calculus for Business" be in "Mathematics" or "Business".

Learners need a set of abstracted common concept rather than instances of existing subjects or curricula for the sake of cross-institute utilization of OER which are provided from various institutes.

2.2 Inflexibleness of Static Hierarchy

Most OER are structured in static hierarchies. A

Takahashi T. and C. Larson R..
AN EFFICIENT EDUCATIONAL MATERIAL EXPLORATION USING EXTRACTED CONCEPTS.
DOI: 10.5220/0003958702820287
In *Proceedings of the 4th International Conference on Computer Supported Education* (CSEDU-2012), pages 282-287
ISBN: 978-989-8565-06-8

learner who aim to complete a course may be able to find an appropriate educational material in the static hierarchies. Learners' objective, however, is not always to complete a course. It may take all sorts depending on learners. while a learner needs to learn whole of elementary calculus, another learner may need only a part of calculus in order to understand Gauss's law in electromagnetism. The existing static hierarchies are not flexible enough to cope with the various learning objectives.

2.3 A Limit of Text Search

There are huge amount of OER now. This makes it difficult for learners to find an appropriate educational material. iTunes University provides a search function. This function looks for educational materials which contain character string given as a query in their subject name or course name. This surface text search is too rough to look for materials regarding learners' precise search demand. And the results tend to be a long list of candidates or empty (so called zero-hit problem).

For instance, as the result of search with a query "Algebra" iTunes University returns a long list consisting of 162 courses which makes learners to consume much time to traverse it. On the other hand, the number of result is zero for a query "Document Classification". Even if the current material collection does not have a suitable course for this topic, some related materials could help learners who want to know about that.

General text search engines return documents which contains query string in them. In order to get suitable result, learners have to know some words or phrases that are supposed to be in desired educational materials, however, learners usually do not have enough knowledge about a new learning object, and it is a difficult task to choose an appropriate word or phrase as a query.

3 OER EXPLORATION USING CONCEPT

We propose to introduce an abstracted expression layer, "concept" which is common to all institutes in order to solve difficulties described in Section 2. The concepts enable the system to manage heterogeneous OER and to provide learners efficient and flexible functions for OER exploration.

Figure 1 shows how the concepts work in the structure. There are educational materials in the bottom layer. Over the materials there are several layers

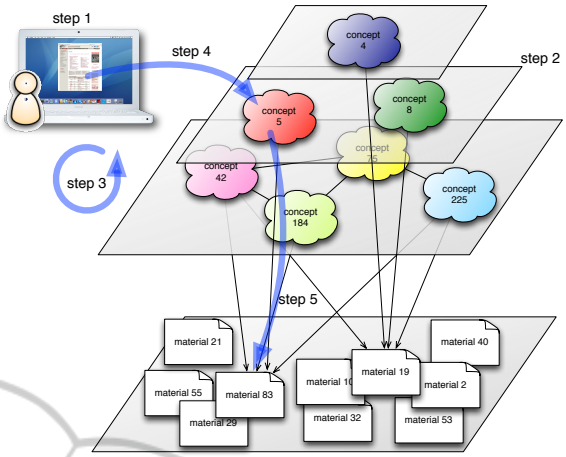

Figure 1: OER management using concepts.

of concept network. Each material has some links to relative concepts. Namely, materials have some concepts as background of the materials, and concepts have some materials which explain the concepts.

Using this structure, system can provide an approach which leads learners to educational materials as the following steps.

1. Learners express their objective or concern by keywords as a search query.
2. System shows (a) a layer of concept network which contains a most related concept to the input keywords and (b) the related materials to each concept.
3. System updates the concept network as learners update the keywords.
4. Learners grasp the target area graphically, and find interesting concept by browsing concept network.
5. System provides appropriate educational materials for the concept.

The concepts are made in a bottom-up manner by the algorithm described in Section 4 instead of a top-down manner that the existing subject or courses are defined in. Because the bottom-up concept does not depend on any specific structure (namely subjects, courses or lectures), it can solve the difficulty of cross-institute, and utilize any OER at the same time.

Since the desired granularity level of concept depends on learners' objectives, we introduced multiple layers for the concept network. Each layer has different granularity of concept which correspond to learners' various levels of objectives. This multiple layers solves the problem of inflexibleness which the existing hierarchies have, and provides learners an efficient search environment. Learners can browse abstracted structures of concept avoiding being buried in a lot of educational materials.

4 CONCEPT EXTRACTION

4.1 Applying LDA to Educational Materials

We used an algorithm Latent Dirichlet Allocation (LDA) proposed by Blei (Blei et al., 2003) to extract concepts from educational materials. LDA is a generative probabilistic model of a document set. The basic idea is that documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words. The model can be represented as Figure 2. In this model, a topic has a word distribution ϕ, and a document has topic distribution θ. LDA gives a set of topics, θ and ϕ given a document set.

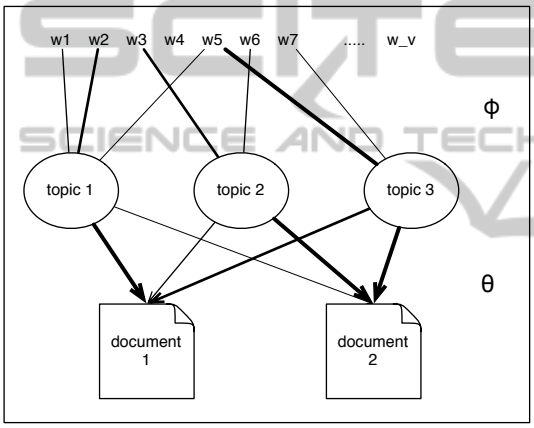

Figure 2: Latent Dirichlet Allocation (LDA).

The structure which LDA is assuming is quite similar to our proposed one shown in Figure 1. Because LDA gives enough data to construct the structure, we chose it for our approach. In the context of OER, the topic and document in LDA can be considered as concept and educational materials respectively. Then by applying LDA, we can get the most likely output described as follows for a given OER.

- a set of concept
- word distribution for each concept (ϕ)
- concept distribution for each material (θ)

Each concept has a word distribution (ϕ) which supports the concept. The distribution enables flexible matching between learners' search query and the concept. Since concept distribution θ reflects the strength of relation between concepts and materials, we can provide relevant materials for a specific concept by using the distribution.

4.2 Layered Concept

LDA is a popular algorithm for the sake of clustering as many researches applied it especially in a research area of Natural Language Processing. The proposed algorithm may be regarded as an application of LDA to a text collection, however, we do not merely aim to make clusters but the followings.

- to find the most appropriate concept for a learners' query
- to find the most relevant educational material for a concept

This motivated us to have multiple sets of clusters instead of only one reasonable set of clusters. A set of clusters forms one concept network layer shown in Figure 1. We proposed to extract multiple layered concept network that allow us to have granularity and flexibility in the matching.

We have to note that each layer does not have any relation with other layers. The layers are made independently, thus the layers shown in Figure 1 do not form a hierarchy.

4.3 Visualization of Concept Network

Since an educational material written in text can be represented as a word vector, the similarity between educational materials can be defined as the similarity of the vectors. On the other hand, a concept have a word distribution ϕ which can be used as a word vector, and the similarity between concepts can be defined as the similarity of the vectors as well.

Once the similarity between elements is defined, the set of elements can be visualized as a network on two-dimensional space like Figure 1 using Force-Directed model proposed by Fruchterman and Reingold (Fruchterman and Reingold, 1991). Force-Directed model assume a physical spring between two nodes. And the position of nodes can be calculated by looking for an equilibrium state in the simulation on dynamic power modeling. We used the similarity described above as strength of the pseudo spring.

The concepts derived by LDA do not have any label. But they have word distribution (ϕ) which can be used to represent a tag cloud.

5 EXPERIMENTAL RESULT

5.1 Experiment

The data we used for our experiment were from two resources, BLOSSOMS (Larson and Murray, 2008)

and MIT OCW. BLOSSOMS had 50 educational video lectures in a variety of topics. Because all videos have a transcript, they are suitable for text processing such as LDA considerably. Though MIT OCW had more than 2,000 lectures, we used 1,019 lectures from which we were able to obtain lecture notes. The proposed algorithm was applied to the transcripts for BLOSSOMS, and text data which were extracted from lecture notes for MIT OCW respectively. While OER are provided in various formats, the proposed approach is applicable to them as long as they have text.

After word regularization such as lemmatize and elimination of stop-words, we applied LDA and derived concepts. We used an algorithm proposed by Griffiths (Griffiths and Steyvers, 2004) for the implementation of LDA. LDA requires several parameters that we set as the following.

- hyper parameter α : 0.5
- hyper parameter β : 0.5
- number of iteration: 100
- number of concept: 10 to 50 step by 5

We used multiple values for the number of concepts as proposed in Section 4.2. By this experiment setting, we obtained nine different layers of concept which were generated independently.

Figure 3 shows an example of output. The system uses the output of LDA to provide search function and to visualize concept network and material network. The probabilistic word distribution ϕ works for scoring of the search function. The score of a concept can be formulated as the summation of ϕ value(s) for query word(s).

The figure "1) Concept Network" in Figure 3 shows the result of the search. For the learner's query, "star", system shows a layer that has 45 concepts in which the concept number 20 is highlighted as the most relevant concept to the query. The summary of the concept number 20 is shown by a tag cloud below the concept network. The learner can browse among other concepts by moving mouse cursor over concept nodes in this network. Browsing concepts, the learner may find a new keyword and add it to the query. "2) Updated Concept Network" shows the result of updating the query by adding "momentum".

If the learner selects the concept number 30 by clicking the node in this network, the system shows a network of educational materials which related to the selected concept as shown in the figure "3) Material Network". The size of nodes in this network reflects the degree of relevance to the selected concept. The learner can grasp a summary of an educational material with a tag cloud by moving mouse cursor over a node in the network as well as the concept network. The system can also show a similarity network in where a selected node and its similar nodes are represented in a network as shown in "4) Similarity Network". Browsing concept network, material network and similarity network, the learner can grasp the target area and select a appropriate material to where the learner can jump, and then start learning with a material as shown in "5) Material Page".

5.2 Evaluation

The result shows that materials from both BLOSSOMS and MIT OCW are used all together. The figure "3) Material Network" in Figure 3 shows the two lectures from different resources ("0315": MIT OCW lecture "Extrasolar Planets: Physics and Detection Techniques", "22": BLOSSOMS video "Galaxies and Dark Matter") are represented in one network. The proposed approach can incorporate any OER provided by different institutes and make concept networks which does not depend on any existing structure. On the concept networks, learners can look for materials which are appropriate to their educational demand. This result shows the algorithm works to solve the first and second problem described in Section 2.

The proposed algorithm gives the interactive search with abstracted concept level in which learners can grasp the target area and related ones. This helps learners to avoid facing a long list of search result. Learners can view OER gradationally from bird's-eye-view to the detail. In the interactive search, the tag clouds over the concept network help learners to find unexpected keywords.

Theoretically, concept networks give extensional expression which explains a target concept by the related materials and the neighbor concepts. The extensional expression works especially in a case that learners are looking for materials to solve their own problem. Because learners usually lack knowledge about the target itself. The experimental result shows that related materials and neighbor concepts can help learners to understand the target. From this result, the proposed approach solved the third difficulty described in Section 2, limit of text search.

6 RELATED WORK

In this work, we aim to provide learners suitable educational materials that relevant to their demand. This objective is the same as what many researches are tackling in an area of information retrieval (Manning

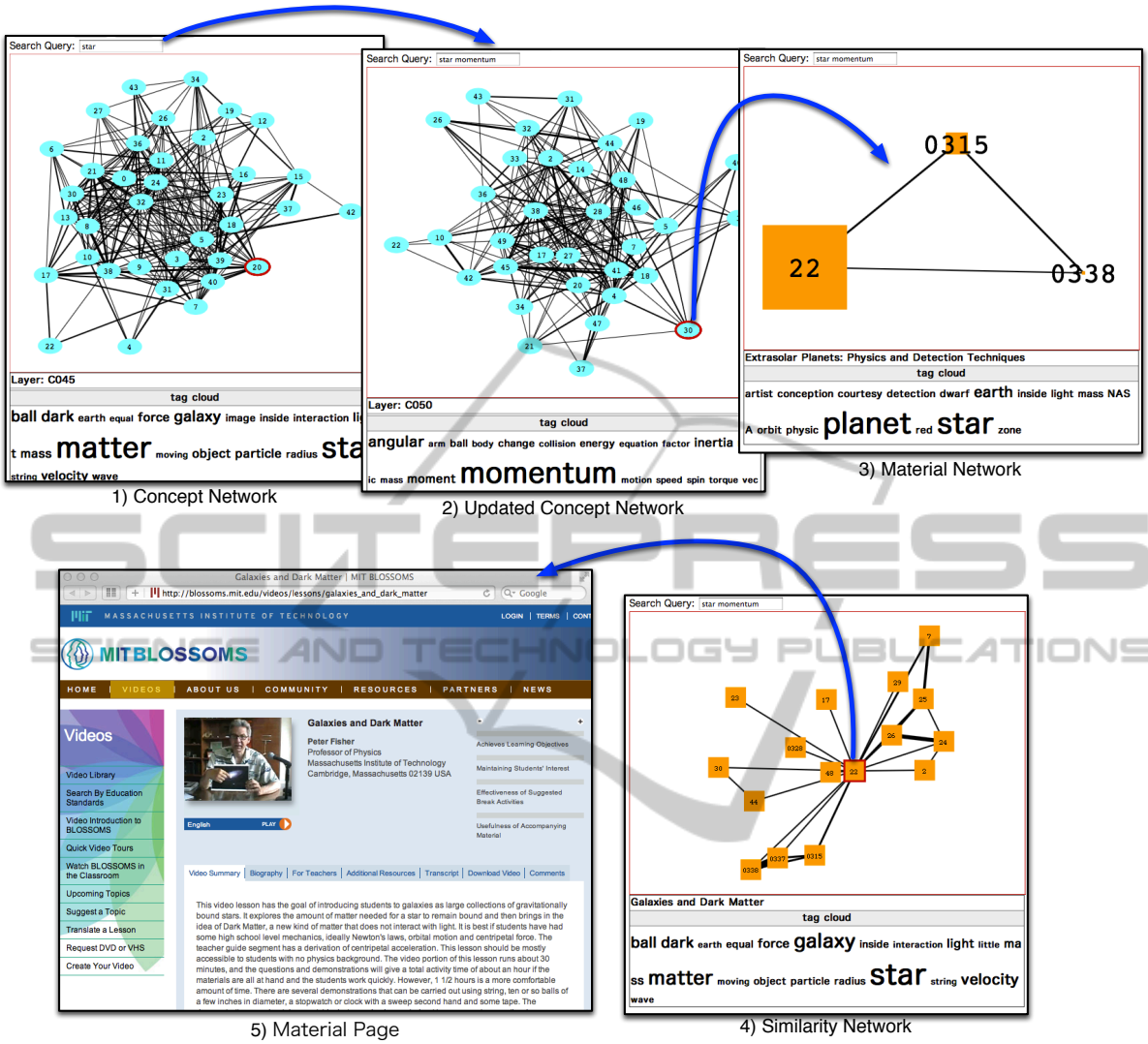

Figure 3: Exploration Steps.

et al., 2008). In most cases of information retrieval, users select a desired content from a list of contents which system generated for a given query. This work flow assumes that users have ability to select desired one from the list. In the context of educational material search, however, learners usually do not have enough knowledge about the target area. Hence it is difficult to select suitable one for the learners. The layered concept network which we proposed helps learners to have a bird's-eye view and to understand the target area. By the interactive interface, learners can find not only what kind of concepts are there for the given keywords, but also what are there as similar or related concept, and what do they need to learn the target concept.

Tudhope (Tudhope et al., 1995) and many works tried to represent search results by visual interface such as two-dimensional network instead of a simple list. These techniques correspond to the representation of the bottom layer of Figure 1 in our work. In addition to the representation, the proposed algorithm has upper layers on the bottom one, and utilize them for the search.

Brusilovsky (Brusilovsky and Rizzo, 2002) and Simko (Simko and Bielikova, 2009) proposed to make a concept network (they called this network "concept map") in order to help learners to understand a target area and to find educational materials. The objective is the same to ours. While Simko made a concept network by using a set of concepts defined by hand, we proposed to derive not only relations but also a set of concepts itself from a bottom-up automatically. Brusilovsky derived both concepts and relations by applying Self Organizing Map to OER.

While they used only one concept network, we used multiple layered concept networks. Because the concepts and educational materials have complex relation each other, a two-dimensional visualization is not enough to represent the complete relation consistently [2], especially, in a case that new educational materials are being added incrementally. We, however, do not have to take care the difficulty, because we use a concept network only to guide learners to appropriate educational materials. This means that we do not have to draw a perfect concept network in our work. We just have to represent subminimal relative relations for guiding learners. For this sake, the statistical distribution produced by LDA works effectively and robustly for various kinds of OER.

7 CONCLUSIONS

We proposed a method to guide learners to appropriate educational materials efficiently using a layered concept networks that is derived from existing OER. Using the layered concept networks, we achieved the objectives of this research, namely to provide learners an environment in which they can understand the target area and explore educational materials efficiently. The experimental result shows the proposed approach works to solve difficulties in utilization of existing OER. The advantages of the approach can help learners, however, the result did not show any quantitative efficiency in searching that is one of important future work.

The proposed algorithm classifies materials into multiple concepts. This classification is based on the types of concepts, however, it says nothing about quality or difficulty of materials. Those information must help learners to chose materials which are appropriate for their demand. This should be added by an algorithm to compliment the concept extraction using LDA.

We propose one directions which is worth for future study, dependence extraction. The proposed method uses only similarity relation between concepts and materials defined as the similarity between vectors. The concepts and materials can have dependence as the meaning of pre-condition. For example, learners have to learn “addition” before “multiplication”. We call the relation dependence, and we can say that multiplication depends on addition for the above example. The dependence will help learners to find their necessary concepts or educational materials. Because the number of combination between concepts and materials is very huge, we have to make an algorithm which works automatically.

[2]We cannot represent even a small graph such as a complete graph which consist of four vertex connected with equivalent strength in two-dimension.

REFERENCES

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022.

Brusilovsky, P. and Rizzo, R. (2002). Map-based horizontal navigation in educational hypertext. In *Proceedings of the thirteenth ACM conference on Hypertext and hypermedia*, HYPERTEXT ’02, pages 1–10, New York, NY, USA. ACM.

Carson, S. (2009). The unwalled garden: growth of the opencourseware consortium, 20012008. *Open Learning: The Journal of Open and Distance Learning*, 24(1):23 – 29.

Fruchterman, T. and Reingold, E. (1991). Graph drawing by force directed placement. *Software Practise and Experience*, 21(11):1129 – 1164.

Griffiths, T. L. and Steyvers, M. (2004). Finding scientific topics. *PNAS*.

Larson, R. and Murray, M. (2008). Open educational resources for blended learning in high schools: Overcoming impediments in developing countries. *Asynchronous Learning Networks*, 12(1).

Manning, C. D., Raghavan, P., and Schuetze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.

Reid, A. (2008). Portable composition: itunes university and networked pedagogies. *Computers and Composition*, 25(1):61 – 78.

Simko, M. and Bielikova, M. (2009). Automatic concept relationships discovery for an adaptive e-course. In *Educational Data Mining*, pages 171–178.

Tudhope, D., Taylor, C., and Benyon-Davies, P. (1995). Navigation via similarity in hypermedia and information retrieval. In *HIM’95*, pages 203–218.