Folksonomy-based Indexing for Location-aware Retrieval of Learning Contents

Wen-Chung Shih[1] and Shian-Shyong Tseng[1,2*]

[1]*Department of Computer Science*
Chiao Tung University, Hsinchu, 30010, Taiwan
{gis90805, sstseng}@cis.nctu.edu.tw

[2]*Department of Information Science and Applications*
Asia University, Taichung, 41354, Taiwan
sstseng@asia.edu.tw

Abstract

With the fast development of wireless communication and sensor technologies, ubiquitous learning has become a promising learning paradigm. In context-aware ubiquitous learning environments, it is desirable that learning content is retrieved according to environmental contexts, such as learners' location. However, traditional information retrieval schemes are not designed for content retrieval in ubiquitous learning environments. Recently, folksonomies have emerged as a successful kind of applications for categorizing web resources in a collaborative manner. This paper focuses on the index creation problem for location-aware learning content retrieval. First, we propose a bottom-up approach to constructing the index according to the similarity between tags, which considers metadata and structural information of the teaching materials annotated by the tags. Then, a maintenance mechanism is designed to efficiently update the index. The index creation method has been implemented, and a synthetic learning object repository has been built to evaluate the proposed approach. Experimental results show that this method can increase precision of retrieval. In addition, impacts of different similarity functions on precision are discussed.

Keywords

Folksonomy; Location-aware; Information retrieval; Ubiquitous learning

1. Introduction

With the fast development of wireless communication and sensor technologies, ubiquitous learning (u-learning) has emerged as a promising learning paradigm, which can sense the situation of learners and provide adaptive supports to students [1] [1-5]. Context-awareness is one major characteristic of u-learning, where the situation or environment of a learner can be sensed. Advantages of context-aware learning are two-folded. In the passive aspect, it can alleviate environmental limitations. In the active aspect, it can utilize available resources to facilitate learning.

There are several types of applications for context-aware u-learning. A typical scenario is "learning with on-line guidance," as presented in [4], which considers the "identification of plants" unit of the Nature Science course in a elementary school. The context is in campus, and the human-system interaction is as follows:

- System: Can you identify the plant in front of you?
- Student: Yes.
- System: What is the name of this plant?
- Student: Ring-cupped oak.
- System: Do you see any insect on it?
- Student: Yes.
- System: Can you identify this insect?
- Student: No.
- …

* Corresponding author

0-7695-3108-3/08 $25.00 © 2008 IEEE
DOI 10.1109/WMUTE.2008.11

The assumption is that the system is aware of the location of the student, and the nearby plants, by sensor technologies and built-in campus maps.

Retrieval of learning content, hereafter named Content Retrieval (CR), is an important activity in u-learning, especially for on-line data searching and cooperative problem solving. Furthermore, both teachers and students need to retrieve learning content for teaching and learning, respectively. Conventional keyword-based content retrieval schemes do not take context information into consideration, so they cannot satisfy the basic requirement of u-learning, which is to provide users with adaptive results. To support context-aware learning, learning content needs to be provided according to learners' contexts. For example, when a student can not identify an insect in the u-learning course, s/he can access a learning object repository for more information by submitting a query. As we can image, queries are most likely ambiguous and need refinement. If context information can be applied to refine the original query, it will be easier for learners to retrieve relevant content.

We classify the schemes of content retrieval into static and dynamic ones according to the adaptability of the retrieved results. For static CR, the retrieved result only depends on the query, independent of users and contexts. Dynamic CR can be further divided into personalized, context-aware and other schemes, according to the factors that are considered by the adaptive mechanisms of CR. Personalized CR is adapted to subjective factors of learners, such as user profile, preference, etc. In other words, the same query submitted by different persons could result in different results retrieved. Context-aware CR is adapted to objective factors of learners, like time, place, device, activity, peers, etc. Hence, the same query issued in different contexts could get different results.

To support context-aware CR, the teaching materials stored in the repositories have to be organized according to their contextual information, in order for efficient retrieval. For example, the relevance of a teaching material about fern plants to a given query depends on the learner's location. In an outdoor learning activity, the desired contents are usually those addressing subject materials nearby. However, it is almost impossible to request experts to annotate all these contents with suitable context-aware metadata. Therefore, content annotation based on folksonomies and automatic techniques in a collaborative way is a promising solution.

In well-known folksonomy applications, such as del.icio.us (http://del.icio.us/), Flickr (http://www.flickr.com), etc., folksonomies are organized into a flat structure, which consists of several catagories named by user-defined tags. This flat organization is suitable for users to manage their preferences. However, when the size of repositories gets larger and larger, a hierarchical organization becomes a better choice to organize the contents. To bridge the gap of the two structures, we formulate this index creation problem and propose a bottom-up approach to constructing an index from existing folksonomies according to the similarity between tags, which considers metadata and structural information of the teaching materials annotated by the tags. Then, a maintenance mechanism is designed to efficiently update the index. The index creation method has been implemented, and a synthetic learning object repository has been built to evaluate the proposed approach. Experimental results show that this method can increase precision of retrieval. In addition, impacts of different similarity functions on precision are discussed.

The contributions can be summarized as follows. First of all, we propose a folksonomy-based method for index creation, which can reduce the effort required in subsequent work of location-aware content retrieval. With this method, the heavy burden of experts for manually developing concept hierarchy can be significantly alleviated. Second, a similarity function is proposed to increase the precision of folksonomy fusion, which considers more characteristics of learning contents, including metadata and structural information of the teaching materials annotated by the tags. Next, a self-organizing mechanism was designed to balance the number of documents annotated by the tags, which can increase the performance of indexing structures. Finally, the proposed method is implemented and the built index is evaluated. Experimental results reveal that this method can improve the performance of retrieval.

2. Preliminaries and Related Work

SCORM is a set of specifications for developing, packaging and delivering high-quality education and training materials whenever and wherever they are needed [6, 7]. In SCORM, content packaging scheme is proposed to package the learning objects into standard teaching materials. Su et al. proposed a level-wise approach to SCORM content management [8]. In this two-phase scheme, structural information is considered. However, the design of this structure doesn't consider its usability in ubiquitous learning environments. There have been numerous studies on Structured Document Retrieval [9, 10]. However,

intra-document structural modeling is not suitable for SCORM-compliant documents.

Inverted file indexing has been widely used in information retrieval [11-14]. An inverted file is used for indexing a document collection to speed up the searching process. However, the structure of a document is not considered in this model. Storage requirements of inverted indices [15] have been evaluated based on B+-tree and posting list. In [16], 11 different implementations of ranking-based text retrieval systems using inverted indices were presented, and their time complexities were also investigated. The meta-search approach has been studied in the context of distributed information retrieval [17]. This approach consists of Query Distribution and Result Merging phases.

We review ontology building approaches because its construction process is similar to index creation in concept. Well-known research approaches to ontology building include [18]: Dictionary-based approach [19], Conceptual clustering [20], Association rules mining [21], Formal concept analysis [18], etc.

3. Problem Formulation

A folksonomy is a user generated taxonomy used to categorize and retrieve web content such as Web pages, photographs and Web links, using user-defined labels called tags. In this work, folksonomies are modeled as follows.

Definition. Folksonomy

A folksonomy defined by a user is a triple of (T, I, R), where T is a set of tags, I is a set of content packages and R is a relation on $T \times I$. We say that tag t is related to a content package i if i is annotated by t. Each tag is associated with two types of attributes, which are derived from content packages annotated by this tag.

- Level j feature vector, L_j, ($0 < j <$ *Height*)

where L_j = the average of the level j feature vectors of content packages annotated by the tag, and *Height* is the height of the rooted tree;

- Metadata, $\{M_k \mid k = 1$ to m, m is the number of metadata$\}$

where the value of M_k is defined by the most frequent M_k value of the content packages annotated by the tag.

Based on the definitions above, the index creation problem is stated as follows.

Definition. The Folksonomy-based Index Creation Problem (FICP)

Given a collection of content packages and corresponding folksonomies, generate an index. The objective is to improve performance of learning content retrieval.

4. Folksonomy-based Index Creation

Our idea to solve this problem is based on the heuristic that existing folksonomies generated by users can be a good starting point from which to construct a location-aware index. While most folksonomies are organized into flat structures, we plan to build hierarchical indices to better organize the learning content. To bridge the gap of the two structures, folksonomies and indices, we propose a bottom-up approach to construct an index from existing folksonomies according to the similarity between tags, which considers metadata and structural information of the teaching materials annotated by the tags.

4.1 The Folksonomy-based Index Creation Algorithm

We design an algorithm to merge two folksonomies into one which is used for subsequent information retrieval. The idea is to make decisions of tag merging according to the similarity of the two tags. The main difficulty is how to choose a suitable similarity function for SCORM-compliant teaching materials, which are characterized by textual content, metadata and structural information. Here, a similarity measure for two tags, a and b, is proposed:

$$Sim(a,b) = (1-\beta)\sum_{i=0}^{Height} \alpha_i \times Sim_i(a,b) + \beta \times Sim_M(a,b)$$

where the sum of α_i is equal to one, and $0<\beta<1$. The parameter, α_i, is used to adjust the weighting of level-wise content vector. The parameter β is used to adjust the weighting of metadata similarity. The similarity function consists of two parts:

- Sim_i: level i similarity function, which is *cosine* function, $0< i <$ *Height*. The similarity between two vectors $v_k = <k_1, k_2, \ldots, k_{|V|}>$ and $v_p = <p_1, p_2, \ldots, p_{|V|}>$ is measured by the following formula:

$$sim_i = \frac{\sum_{i=1}^{|V|} k_i \times p_i}{\sqrt{\sum_{i=1}^{|V|} {k_i}^2} \times \sqrt{\sum_{i=1}^{|V|} {p_i}^2}}$$

- Sim_M: Metadata similarity function, which is (the number of matched attributes) / (the number of all attributes).

The Folksonomy-based Index Creation Algorithm is listed as follows.

Algorithm 1. Folksonomy-based Index Creation Algorithm (AlgFIC)
Input:
F_1, F_2: the two folksonomies to be merged
Th_sim: the threshold for comparing similarity
sim: the similarity function
Output:
F: the merged folksonomy

Step 1. Initialization
1.1 Each tag of F_1 and F_2 is represented by the average of its related teaching materials.
1.2 F_1 is assigned as the Master folksonomy, M.
Step 2. for each tag t in F_2
2.1 calculate the similarity of t and each tag of M.
2.2 let *t_close* be the closest tag in M to t.
2.3 if the $sim(t, t\_close) > Th\_sim$ then
add the tag t into *t_close*
else
add tag t into M as a new tag

5. Implementation

A folksonomy-based information retrieval system is implemented to evaluate the proposed method. To evaluate the performance of the proposed approach for information retrieval, three experiments are conducted. Two synthetic learning object repositories (LOR) are used in this experiment. The first LOR contains 1,200,000 SCORM-compliant documents [22], which are converted from Web pages related to educational domains. The other LOR contains 2,400,000 SCORM-compliant documents, which are converted from technical papers related to computer science domains.

We compare the performance of the proposed method with that of the keyword-based method. The values of α_1 and α_2 are both 0.5. The value of β is 0.3. As shown in Figure 1, the proposed method can significantly improve the performance with respect to precision and recall.

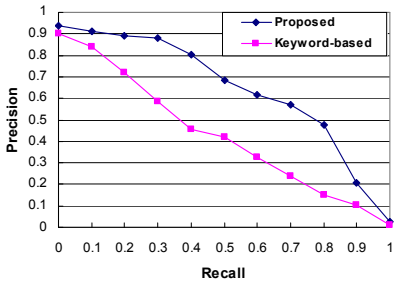

Figure 1. Comparison with the keyword-based method

We compare the performance of different similarity functions. As shown in Figure 2, the proposed similarity (All) got the best performance. Also, the level-wise similarities have larger impact on performance than metadata. Finally, the level_1 similarity has large impact than Level_2.

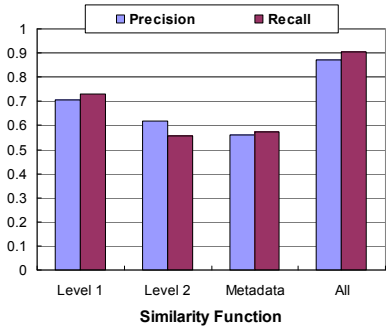

Figure 2. Comparison of Different Similarity Functions

The results show that the proposed approach attains better performance than the traditional keyword-based search. The primary reason is that the proposed method uses the location-aware index to conduct semantic search, trying to find various semantic meanings of a given query. For example, a query for "fern plants" would return relevant results about plants which are present nearby, in a location-aware manner. Therefore, the proposed method can get better precision.

6. Conclusions

This paper describes a folksonomy-based approach to index creation for location-aware learning content retrieval. We divide the index creation problem into two sub-problems: folksonomy fusion and hierarchy organizing. The former problem is solved by merging tags according to a similarity function, which considers textual, metadata and structural information of teaching materials. Also, we design a self-organizing mechanism to address the latter problem. Experimental results show the effectiveness of this approach. This folksonomy-based approach is characterized by a time-saving development process, minimal involvement of experts and high performance of information retrieval.

Acknowledgements

This research was partially supported by Science Council under the

number of NSC95-2520-S009-007-MY3 and NSC95-2520-S009-008-MY3.

References

[1] C.-M. Chen, Y.-L. Li, and M.-C. Chen, "Personalized Context-Aware Ubiquitous Learning System for Supporting Effectively English Vocabulary Learning," in *Advanced Learning Technologies, 2007. ICALT 2007. Seventh IEEE International Conference on*, 2007, pp. 628-630.

[2] S. J. H. Yang, "Context Aware Ubiquitous Learning Environments for Peer-to-Peer Collaborative Learning," *Educational Technology & Society,* vol. 9, pp. 188-201, 2006.

[3] N.-K. Si, J.-F. Weng, and S.-S. Tseng, "Building a Frame-Based Interaction and Learning Model for U-Learning," in *Ubiquitous Intelligence and Computing*, 2006, pp. 796-805.

[4] G.-J. Hwang, "Criteria and Strategies of Ubiquitous Learning," in *Sensor Networks, Ubiquitous, and Trustworthy Computing, 2006. IEEE International Conference on*, 2006, pp. 72-77.

[5] F.-R. Kuo, G.-J. Hwang, Y.-J. Chen, and S.-L. Wang, "Standards and Tools for Context-Aware Ubiquitous Learning," in *Advanced Learning Technologies, 2007. ICALT 2007. Seventh IEEE International Conference on*, 2007, pp. 704-705.

[6] E. D. Nitto, L. Mainetti, M. Monga, L. Sbattella, and R. Tedesco, "Supporting Interoperability and Reusability of Learning Objects: The Virtual Campus Approach," *Educational Technology & Society,* vol. 9, pp. 33-50, 2006.

[7] J.-M. Su, S.-S. Tseng, W. Wang, J.-F. Weng, J. T. D. Yang, and W.-N. Tsai, "Learning Portfolio Analysis and Mining for SCORM Compliant Environment," *Educational Technology & Society,* vol. 9, pp. 262-275, 2006.

[8] J. M. Su, S. S. Tseng, C. Y. Wang, Y. C. Lei, Y. C. Sung, and W. N. Tsai, "A Content Management Scheme in SCORM Compliant Learning Object Repository," *Journal of Information Science and Engineering,* vol. 21, pp. 1053-1075, 2005.

[9] A. Trotman, "Searching structured documents," *Information Processing and Management,* vol. 40, pp. 619-632, 2004.

[10] A. Trotman, "Choosing Document Structure Weights," *Information Processing and Management,* vol. 41, pp. 243-264, 2005.

[11] R. Baeza-Yates and B. Ribeiro-Neto, *Modern information retrieval*. New York: ACM Press, 1999.

[12] A. Mittal, P. V. Krishnan, and E. Altman, "Content Classification and Context-Based Retrieval System for E-Learning," *Journal of Educational Technology & Society,* vol. 9, pp. 349-358, 2006.

[13] G. Salton and M. J. McGill, *Introduction to Modern Information Retrieval*. New York: McGraw & Hill, 1983.

[14] I. H. Witten, A. Moffat, and T. C. Bell, *Managing gigabytes: compressing and indexing documents and images*, 2nd ed. ed. San Francisco, California: Morgan Kaufmann, 1999.

[15] Y. K. Lee, S.-J. Yoo, K. Yoon, and P. B. Berra, "Index structures for structured documents," in *the 1st ACM international conference on digital libraries*, 1996, pp. 91-99.

[16] B. B. Cambazoglu and C. Aykanat, "Performance of query processing implementations in ranking-based text retrieval systems using inverted indices," *Information Processing and Management,* vol. 42, pp. 875-898, 2006.

[17] C. Yu, K. L. Liu, W. Meng, Z. Wu, and N. Rishe, "A Methodology to Retrieve Text Documents from Multiple Databases," *IEEE Transactions on Knowledge and Data Engineering,* vol. 14, pp. 1347-1361, 2002.

[18] S.-S. Weng, H.-J. Tsai, S.-C. Liu, and C.-H. Hsu, "Ontology construction for information classification," *Expert Systems With Applications,* vol. 31, pp. 1-12, 2006.

[19] L. Khan and F. Luo, "Ontology Construction for Information Selection," in *the 14th IEEE internation conference on tools with artificial intelligence*, Washington DC, 2002, pp. 122-127.

[20] A. Hotho, A. Maedche, and S. Staab, "Ontology-based Text Clustering," in *the IJCAI-2001 workshop text learning: Beyond supervision*, Seattle, 2001.

[21] A. Maedche and S. Staab, "Ontology Learning for the Semantiv Web," *IEEE Intelligent Systems,* vol. 16, pp. 72-79, 2001.

[22] SCORM, "Sharable Content Object Reference Model (SCORM)." vol. 2006: Advanced Distributed Learning, 2004.