Towards open corpus adaptive hypermedia: a study of novelty detection approaches

Yi-ling Lin and Peter Brusilovsky

School of Information Sciences, University of Pittsburgh,
135 North Bellefield Avenue, Pittsburgh, PA 15206, USA
{yil54,peterb}@pitt.edu

Abstract. Classic adaptive hypermedia systems are able to track a user's knowledge of the subject and use it to evaluate the novelty and difficulty of content encountered by the user. Our goal is to implement this functionality in an open corpus context where a domain model is not available nor is the content indexed with domain concepts. We examine methods for novelty measurement based on automatic text analysis. To compare these methods, we use an evaluation approach based on knowledge encapsulated in the structure of a textbook. Our study shows that a knowledge accumulation method adopted from the domain of intelligent tutoring systems offers a more meaningful novelty measurement than methods adapted from the area of personalized information retrieval.

Keywords: Novelty detection, knowledge modeling, personalization

1 Introduction

The World Wide Web greatly increased the volume and the variety of educational content available to the public. However, the abundance of content makes it difficult for users to find "the right content" that matches their individual goals, interests, and knowledge level. A user may benefit from personalized guidance to help manage and navigate through this abundance of resources. In a number of educational adaptive hypermedia (AH) systems, adaptive navigation support techniques were able to help individual students locate, recognize, and comprehend relevant information, thus increasing learning outcomes and retention [1],[2],[3],[4]. Unfortunately, these systems cannot be directly applied to an open corpus of Web educational content. Existing adaptive navigation support techniques are only able to work within a closed corpus of documents that have been manually structured and indexed with domain concepts and metadata at design time; however, they are impractical for most web-based real world applications.

We believe that the field of educational AH has to undergo the same transformation as the field of information retrieval (IR) did when it moved from manual indexing to automatic indexing procedures. User modeling and adaptation techniques based on manual concept indexing must be augmented or replaced by techniques based on automated text analysis (ATA). This transformation

will make it possible to provide personalized educational guidance for large volumes of online content. Some approaches to building open corpus hypermedia using ATA have been already explored [5]. This study takes another step towards open corpus educational AH by exploring several ATA-based approaches to knowledge-based novelty detection.

2 Novelty Detection for Educational AH

Let us imagine the common situation where students are studying a specific concept in a class. If the topic is one of the main concepts in the class, relevant content can be found in several different textbooks and online sources. Suppose a student has read a textbook section devoted to the target concept and wants more information about it. Ideally, the very next section will offer more information about the concept. The author of this textbook assumes that the new content is suited to the student's already acquired knowledge. However, this assumption doesn't hold if the new content is found in another textbook or on a Web page. While search engines might help the student to find dozens of pages with relevant contents, no search engine can ensure that this content is suitable for the student's knowledge level. Pages that are "just right" (new, and ready to be learned) will be intermixed with multiple pages that present learned information about a concept in a varied way and pages with new content at yet a more complex level than the student is capable of understanding.

Adaptive navigation support in classic educational AH was able to warn the user about "nothing new" and "not ready" pages [4], however, it was based on manual page indexing with concepts. In our research, we attempted to recreate a part of this functionality by developing an ATA-based approach to knowledge-based novelty detection. This approach aims to provide open corpus with an adaptive navigation support which can warn users about pages that might have little or no new content and distinguish them from pages that present new content. The open corpus version of the "not ready" approach is not specifically considered in this paper; however, one can argue that the ability to find pages with a very high level of novelty could be the closest analogy to the "not ready" functionality in classic AH which can warn the student of too advanced content. To achieve the goal we explored three straightforward approaches to novelty: vector space approach, language model approach, and knowledge modeling approach.

Vector Space Model approach is based on the classic IR algebraic model for representing text documents [6] which is commonly used for IR user profiling approaches. Each document is represented as a vector in m-dimensional space using TD-IDF as the weighting scheme. The fundamental intuition of TF-IDF is a) the more frequent the term is, the more indicative the term is of the topic, and b) the less frequent the term is in the corpus, the greater power the term could have to discriminate the importance of the term in the corpus. The document is denoted as a vector $d_i = (w_1(d_i), w_2(d_i), ..w_m(d_i))^t$.To represent a student's knowledge, we used the centroid of documents viewed by the student.

If the student has read $d_1, d_2, ...d_n$, the student knowledge vector could be represented as $d_i = (\sum_{k=1}^{n} \frac{w_1(d_k)}{n}, \sum_{k=1}^{n} \frac{w_2(d_k)}{n}, ..., \sum_{k=1}^{n} \frac{w_m(d_k)}{n})$. In this context, document novelty is a measure of dissimilarity between document vector and student knowledge vector. Since cosine similarity is the standard similarity measure in IR context [7], we define one minus cosine similarity of these vectors as our measure of novelty.

Language Model is a probabilistic distribution that captures the probability of a sequence of features. In modern IR, it has shown promise for identifying relevant documents in different tasks [8],[9],[10]. A natural approach to novelty detection using a language modeling approach is estimating the likelihood that a set of documents viewed by a certain student and an upcoming new document are generated by the same language model. Kullback-Leibler (KL) divergence is a distributional similarity measure to estimate the redundancy of one document d given a set of viewed document. $R(d_t|d_i) = -KL(\theta_{d_t}, \theta_{d_i}) = -\sum_{w_i} p(w_i|\theta_{d_i}) \log(\frac{p(w_i|\theta_{d_i})}{p(w_i|\theta_{d_i})})$. In the language model approach, a document d is represented by a unigram word distribution θ_d, and it is a multinominal distribution. θ_d can be simply estimated by maximum likelihood estimation (MLE). The problem with using MLE is that it will get a zero probability if a word never occurs in a document d. If a word is in d_t but not in d_i, it will cause $KL(\theta_i|\theta_j) = \infty$. The Dirichlet distribution [10] is a smoothing technique using the conjugate prior for a multinominal distribution. It could be used to adjust the amount of reliance on the words according to the total number of the words. For a Dirichlet distribution with parameter $(\lambda p(w_1), \lambda p(w_2), ...\lambda p(w_m))$. The posterior distribution using Bayesian analysis is $P_\lambda(w_i|d) = tf(w_i, d) + \lambda p(w_i)/\sum_{w_j}(tf(w_j, d) + \lambda p(w_j))$

Knowledge Model approach is our attempt to implement classic knowledge modeling from the domain of intelligent tutoring systems in the open corpus context. This classic approach is based on concept-level or skill-level domain models and uses an overlay model of user knowledge that measures the probability that the user knows a concept or has mastered a skill. For our model, we used the Bush-Mosteller-Atkinson asymptotic modeling approach [11] replacing traditional concept with words extracted by ATA. The knowledge K of each word in a student knowledge vector is: $K_0 = 0, K_{n+1} = K_n + pV \times \frac{W_{i,d_{n+1}}}{\sum_i W_{i,d_{n+1}}}$ where: pV is the speed of knowledge growth for a student (ranged from 0 to 1 and set as an average 0.5 in our experiments). $W_{i,d_{n+1}}$: the weight of word i in document d_{n+1} which is the most recent document. $\sum_i W_{i,d_{n+1}}$: the sum of all word weights in document d_{n+1}. A new document could be represented as a vector: $d_i = (pV \times \frac{W_{1,d_i}}{\sum_i W_{i,d_i}}, pV \times \frac{W_{2,d_i}}{\sum_i W_{i,d_i}}, ...pV \times \frac{W_{m,d_i}}{\sum_i W_{i,d_i}})^t$. This knowledge modeling approach replaces the IR-based centroid model of the vector space approach, retaining cosine approach to novelty calculation.

3 Experimental Methodology

A proper evaluation of a novelty approach is a challenging task that requires a large-scale user study. We believe, however, that a meaningful comparison of

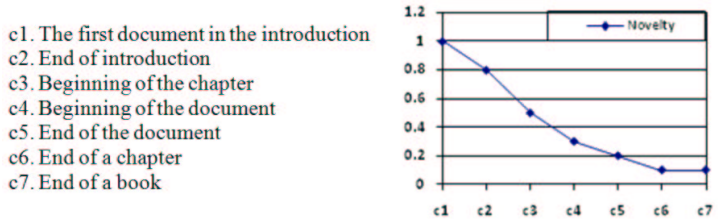

Fig. 1. The novelty trend.

novelty approaches can be performed using an expert writer's knowledge encapsulated in the structure of a textbook.

Our idea is based on a previously-mentioned assumption that a good textbook is constructed from sequentially-written chapters where each new subsection has a reasonably stable level of novelty from the perspective of a sequential reader. Consider a subsection in the middle of a chapter (for example, 10.2.1). At the beginning of a book, the subsection should be completely new to the student (highest novelty value). After reading the introduction, the student should have a general idea about book topics, which will reduce the novelty value for the subsection. Likewise, after reading the introduction to chapter 10, the student will have an even better idea of the contents of that chapter which will decrease the novelty value for the remaining documents in the chapter. After reading the probed document (10.2.1), its novelty value should drop to a very low level. After that, the novelty should change very little, although we assume that a minimum novelty remains until the end of the book. Our method simply evaluates how a specific model represents an expected decrease in document novelty by examining several criteria shown in Fig. 1.

4 Evaluation

In our study, we use the textbook, **Interactive System Design**[12], containing 15 chapters, 504 pages, and 399 documents (numbered subsections). We examined the novelty trend for all the documents in each chapter with our three models. A sample of this analysis for three different chapters is shown in Fig. 2. Our assumption was that a better novelty measurement approach should more closely model the expected declining novelty trend. We also assumed that, in a better model, the novelty decrease trajectories of documents in the same chapter should be reasonably similar to each other due to their comparable amount of novelty and position in the book.

We analyze this prospect the novelty trajectories of the vector space model shown in Fig. 2a. The left graph represents the novelty of 18 documents in chapter 3 computed for each checkpoint. As we can see, the trajectories do not match our expectation. The novelty of documents remains relatively high; even after the chapter has been read. Moreover, the expected decrease in novelty

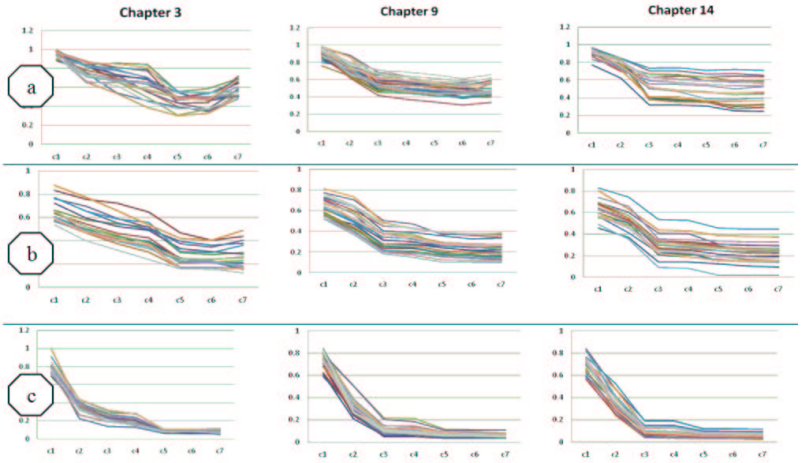

Fig. 2. (a)The novelty trend of vector space model; (b) The novelty trend of language model (c) The novelty trend of knowledge model.

clearly changes to a counterintuitive increase after c6, which is observed for all chapters except 13 & 14. In addition, document novelty trajectories greatly differ within the same chapter.

For the language model, in order to make it comparable with other models, the KL values have been normalized by the maximum of the KL results. In our experiment, the larger the number for divergence, the more novel it is. The behavior of this approach (Fig. 2b) is closer to the expected than the previous one. Yet, we still observe unnatural novelty increases and the novelty trajectories in this model are more spread out. We suspect that the model might be more sensitive inorder to predict the probability of each word in the document. It is better at identifying a topic from a document, but not as good in identifying the novelty of the document from a set of documents read by the student.

The knowledge model produces more consistent trajectories than the other two models and the pattern is closer to our expectations (Fig. 2c). The only unexpected trend is a too steep drop at the beginning of the pattern; however, the rate of decrease depends of the learning speed and can be matched to the expected behavior by selecting the proper speed. In the future, we will have further studies on those factors.The results delivered by our evaluation process should not be considered as a proof that the knowledge model provides a reliable mechanism for novelty modeling. The study was designed not to prove the quality of a specific approach (for that we would need real users and a much larger variety of content than sections from the same book), but to forecast which of the three roads is more promising for further work on novelty detection in an education context. Contrary to the current trends in novelty detection, which are solely focused on IR approaches, our study indicates that a combination of IR document processing with knowledge modeling might be more promising.

5 Conclusion

Our paper attempts to contribute to solving the open corpus adaptive hypermedia problem by comparing several novelty measurement approaches based on ATA. Using an original evaluation method based on knowledge encapsulated in a textbook structure, we compared two approaches inspired by classic and modern information retrieval ideas with an approach inspired by intelligent tutoring ideas. Our results indicate that traditional IR modeling approaches that are known to work well for interest modeling might not be appropriate for knowledge modeling and novelty estimation. In contrast, knowledge modeling based on the fusion of IR and intelligent tutoring ideas looks promising and has to be investigated further.

References

1. Brusilovsky,P.: Adaptive Hypermedia. User Modeling and User Adapted Interaction. 11(1/2), 87–110 (2001)
2. Brusilovsky,P., Pesin,L.: Adaptive Navigation Support in Educational Hypermedia: An Evaluation of The ISIS-Tutor. Journal of Computing and Information Technology. 6(1), 27–38 (1998)
3. Weber,G., Specht, M.: User Modeling and Adaptive Navigation Support in WWW-based Tutoring Systems. In: 6th International Conference on User Modeling, pp.289–300. Springer Wien, New York (1997)
4. Brusilovsky,P., Eklund, J.: A Study of User-Model Based Link Annotation in Educational Hypermedia. Journal of University Computer Science. 4(4), 429–448 (1998)
5. Brusilovsky,P., Henze, N. eds: Open Corpus Adaptive Educational Hypermedia. In: Brusilovsky, P., Kobsa, A., Neidl, W. (eds.) The Adaptive Web:Methods and Strategies of Web Personalization. LNCS, vol. 4321, pp. 671–696. Springer, Verlag (2007)
6. Salton,G., Wong, A., Yang, C.S.: A Vector Space Model for Automatic Indexing. Communications of the ACM. 18(11), 613–620 (1975)
7. Jones,W.P., Furnas, G.W.: Pictures of Relevance. Journal of the American Society for Information Science. (1987)
8. Kraaij, W., Pohlmann, R., Hiemstra, D.: Twenty-one at TREC-8: Using Language Technology for Information Retrieval. In: 8th Text REtrieval Conference(TREC-8)(1999)
9. Miller, D.R.H., Leek, T., Schwartz, R.: A Hidden Markov Model Information Retrieval System. In: 22th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,pp. 214–221. ACM, New York (2001)
10. Zhai, C., Lafferty, J.: Model-based Feedback in the Language Modeling Approach to Information Retrieval. In: 10th International Conference on Information and Knowledge Management ,pp. 403–410. ACM, New York (2001)
11. Atkinson, R.C., Bower, G.H., Crothers, E.J.: An Introduction to Mathematical Learning Theory. John Wiley & Sons, New York-London-Sydney(1965)
12. Newman, W.M., Lamming, M.G.: Interactive System Design. Addison-Wesley(1995)