Information Retrieval for Language Tutoring: An Overview of the REAP Project

Kevyn Collins-Thompson Jamie Callan

Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213-8213

{kct,callan}@cs.cmu.edu

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Retrieval models.

General Terms

Algorithms, Design, Human Factors.

Keywords

Information retrieval, computer-assisted learning.

1. INTRODUCTION

Typical Web search engines are designed to run short queries against a huge collection of hyperlinked documents quickly and cheaply, and are often tuned for the types of queries people submit most often [2]. Many other types of applications exist for which large, open collections like the Web would be a valuable resource. However, these applications may require much more advanced support from information retrieval technology than is currently available. In particular, an application may have to describe more complex information needs, with a varied set of properties and data models, including aspects of the user's context and goals.

In this paper we present an overview of one such application, the REAP project, whose main purpose is to provide reader-specific practice for improved reading comprehension. (REAP stands for REAder-specific Practice.) A key component of REAP is an advanced search model that can find documents satisfying a set of diverse and possibly complex lexical constraints, including a passage's topic, reading level (e.g. 3rd grade), use of syntax (simple vs. complex sentence structures), and vocabulary that is known or unknown to the student. Searching is performed on a database of documents automatically gathered from the Web which have been analyzed and annotated with a rich set of linguistic metadata. The Web is a potentially valuable resource for providing reading material of interest to the student because of its extent, variety, and currency for popular topics.

2. SYSTEM DESCRIPTION

Here we describe the high-level design of the REAP information retrieval system, including document database requirements and construction, annotations, and a brief description of the retrieval model.

SIGIR '04, July 25-29, 2004, Sheffield, South Yorkshire, UK.
ACM 1-58113-881-4/04/0007

2.1 Database Construction

Our goal is to present passages that are interesting to students, whether they are on current topics such as pop stars or sports events, or related to particular classroom projects. To this end, we use the Web as our source of reading practice materials because of its extent, variety, and currency of information.

We want coverage of topics in the database to be deeper in areas that are more likely to be of interest to students. Coverage of other areas is intended to be broad, but more shallow. We therefore gather documents for the database using focused crawling [3]. The current prototype uses a page's reading difficulty to set priority for all links from that page equally, based on the distance from the target reading level range. We plan to explore more refined use of annotations to direct the crawl on a link-by-link basis. In our prototype, we collected 5 million pages based on an initial set of 20,000 seed pages acquired from the Google Kids Directory [7]. Our goal is to have at least 20 million pages that focus on material for grades 1 through 8. The document database must be large enough that the most important lexical constraints are satisfied by at least a small number of pages. Data annotation is currently performed off-line at indexing time. The specific annotations for REAP are described in Section 2.2.

Once the documents are acquired, they are indexed using an extended version of the Lemur IR Toolkit [9]. We chose Lemur because of its support for language model-based retrieval, its extensibility, and its support for incremental indexing, which is important for efficient updates to the database. Annotations are currently stored as Lemur properties, but later versions will take advantage of the enhancements planned for support of rich document structure, described in Section 2.3.

2.2 Linguistic Annotations

In addition to the underlying text, the following linguistic annotations are specified as features to be indexed:

- Basic text difficulty within a document section or region. This is calculated using a new method based on a mixture of language models [4] that is more reliable for Web pages and other non-traditional documents than typical reading difficulty measures.
- Grammatical structure. This includes part-of-speech tags for individual words as well as higher-level parse structures, up to sentence level.
- Document-level attributes such as title, metadata keywords, and ratings.

- Topic category. This would involve broad categories such as fiction/non-fiction [5] or specific topics, perhaps based on Open Directory.
- Named entity tags. We use BBN's Identifinder [1] for high-precision tagging of proper names.

We may also look at more advanced attributes such as text coherence and cohesion [6].

2.3 Query and Retrieval Models

A typical information need for the REAP system might be described as follows:

> Find a Web page about soccer, in American English, with reading difficulty around the Grade 3 level. The text should use both passive- and active-voice sentence constructions and should introduce about 10% new vocabulary relative to the student's known-vocabulary model. The page's topic is less important than finding pages that practice the words: for example, an article on another sport that satisfies the other constraints would also be acceptable.

Information needs in REAP will be modeled as mixtures of multiple word histograms, representing different sources of evidence, as well as document-level or passage-level constraints on attributes such as reading difficulty. There is precedent for using word histograms to specify information needs: indeed, query expansion is one example of this. More specifically, related work includes language model-based techniques such as relevance models [8].

No current Web-based search engine is able to make use of combinations of lexical constraints and language models in this way, on such a large scale. To support this, we are making extensions to Lemur that include:

1. Retrieval models for rich document structure, which includes nested fields of different datatypes where each field may be associated with its own language model.
2. More detailed retrieval models in which we skew language models towards the appropriate grade level, topic, or style.
3. The use of user model descriptions as context for a query.

2.4 User Profiles

In the current prototype, we model a reader's topic interests, reading level, and vocabulary acquisition goals using simple language models. For example, we model the curriculum as a word histogram. Although crude, this captures word-frequency information associated with general reading difficulty, as well as capturing topics that are the focus of the curriculum at each grade level. We plan to add more complex aspects to user profiles, including more specific lexical constraints such as grammar constructs and text novelty. The models can be updated incrementally as the student's interests evolve and they make progress through the curriculum.

3. EVALUATION METHODS

Evaluation of the end-to-end REAP system will be via a series of three year-long studies with both adults and children. The adult studies will provide feedback on vocabulary matching and comprehension, and the child studies will test the hypothesis that children will read adaptively to texts that vary in vocabulary demands, where those texts that closely reflect the reader's interests and comprehension can be used to support improved comprehension and vocabulary growth.

4. CONCLUSION

The REAP project is intended to advance the state of the art in information retrieval, as well as research in reading comprehension, by bringing together practical user models of student interests, vocabulary knowledge and growth, and other aspects of reading, with interesting material from large, open collections like the World Wide Web. This type of system is a valuable new research tool for educational psychologists and learning scientists, because it gives much greater control over how instructional materials are selected. This in turn allows testing of instructional hypotheses, such as the effect of 10% vocabulary stretch, which have been impractical to test in the past. The work also has direct application to other areas of language learning, such as English as a Second Language training. More broadly, however, we believe the REAP project is a important first step toward enabling richer user and task models than currently available with ad-hoc search systems.

5. ACKNOWLEDGMENTS

We thank our collaborators Maxine Eskenazi, Charles Perfetti and Jonathan Brown; John Cho and Alexandros Ntoulas of UCLA for their crawler code; and the anonymous reviewers. This work was supported by U.S. Dept. of Education grant R305G03123. Any opinions, findings, conclusions, or recommendations expressed in this material are the authors' and do not necessarily reflect those of the sponsors.

6. REFERENCES

[1] Bikel, D. M., Miller, S., Schwartz, R., Weischedel, R. M., Nymbol: A high-performance learning name-finder. In *Proceedings of the 5th Conference on Applied Natural Language Processing*, 194 - 201, 1997.

[2] Broder, A. A taxonomy of Web search. In *SIGIR Forum*, 36(2). 3 - 10, 2002.

[3] Chakrabarti, S., van der Berg, M., & Dom, B. Focused crawling: a new approach to topic-specific web resource discovery. In *Proc. of the 8th International World-Wide Web Conference (WWW8)*, 1999.

[4] Collins-Thompson, K., & Callan, J. A language modeling approach to predicting reading difficulty. *Proceedings of HLT/NAACL 2004*, Boston, USA, 2004.

[5] Finn, A., Kushmerick, N. & Smyth, B. Fact or fiction: Content classification for digital libraries. *Joint DELOS-NSF Workshop on Personalisation and Recommender Systems in Digital Libraries* (Dublin), 2001.

[6] Foltz, P. W., Kintsch, W., Landauer, T. K. Analysis of text coherence using Latent Semantic Analysis. *Discourse Processes* 25(2-3), 285 - 307, 1998.

[7] Google Kids Directory. http://directory.google.com/Top/Kids_and_Teens/

[8] Lavrenko, V., and Croft, B. Relevance-based language models. In *Proc. of the 24th Annual International ACM SIGIR Conference*, New Orleans, 120 - 127, 2001.

[9] Ogilvie, P. and Callan, J. Experiments using the Lemur Toolkit. In *Proc.of the 10th Text Retrieval Conference*, *TREC 2001*. NIST Special Publication 500-250, 103-108, 2001.