SourceFinder: Course preparation via linguistically targeted web search

Irvin R. Katz and Malcolm I. Bauer
Educational Testing Service
Princeton, NJ 08541 USA
Tel: +1 609 734 5150
Fax: +1 609 734 1090
ikatz@ets.org
mmbauer@ets.org

ABSTRACT
The use of the Internet for course preparation is ill served by traditional, content-based search engines. This paper describes SourceFinder, a web search engine that locates text material based on linguistic characteristics, such as reading level. Combining SourceFinder with content-based searches may allow instructors more easily to identify material relevant for their courses. We present examples from the teaching of reading comprehension, history, and statistics of how SourceFinder might aid instructors’ use of the Internet for course preparation.

Keywords
Teaching, Text analysis, Internet search, Linguistic features, Course preparation

The Problem
Course preparation is a difficult, time-consuming process. It is not unusual to hear of new assistant professors spending a day or more of preparation for each hour of class time. At the elementary and secondary school level, even more experienced teachers report spending approximately 5 hours per week in curriculum or lesson planning for their 15 hours per week of actual teaching (TIMSS, 1995). Course preparation can be challenging for experienced instructors because of the need to keep their instructional material current with the field they are teaching and relevant to students’ everyday lives.

In the past few years, instructors have turned to the Internet to aid course preparation. Instead of looking through published textbooks or reviewing test banks produced by publishers, the Internet offers a wider variety of sample instructional material as well as real-world material that can be adapted for instructional use. Thus, to construct lecture notes, tests, and other instructional material, teachers might conduct web searches to find raw materials. While this use of the Internet by instructors is more prevalent, it is ill served by current content-based web search engines, which rely on the information retrieval skills of the instructor. Furthermore, content-based web searches typically return a virtual mountain of information, resulting in the “information overload” problem (Nielsen, 1995): how to find the material most relevant to your need without being overwhelmed by the irrelevant (but related) information generated through content-based (keyword) searches.

Sites such as the WWW Virtual Library are a helpful alternative to general content-based web searches, but are limited to material already identified and organized by others. Furthermore, while such sites might have material in the correct content area, the material is of uneven quality and would take some time for an instructor to even identify instructional material of an appropriate difficulty level for his or her students.

This paper describes SourceFinder, a domain-independent search engine that locates text material based on linguistic characteristics, rather than purely content. Instead of searching based on content, SourceFinder searches a website (following links to a user-specified depth) to locate passages of text that meet characteristics such as a particular reading level, a certain density of argument, and an internally coherent clarity of expression (i.e., the passage can be understood with minimal background knowledge). Such factors cut across content categories, which is needed for applications including (a) constructivist learning approaches that emphasize real-world contexts for student learning activities and assessments and (b) assessments of linguistic competence, which require challenging prompts that nevertheless do not allow content knowledge to help or hinder performance. SourceFinder searches can be combined with more traditional content-based searches by asking SourceFinder to follow the links resulting from a web search engine, such as Yahoo or Google.

The software framework and general approach of SourceFinder has direct applications to preparation for classroom instruction. For example, a writing instructor might use SourceFinder to locate material to serve as
prompts for research paper assignments. Other example applications include preparing complex learning tasks for mathematics students to allow students to apply what they’ve learned to more general contexts, the construction of compelling, real-world examples to illustrate critical points in a class, the identification of material for assessments of reading comprehension, and the development of prompts for authentic assessments.

System Description

Figure 1 depicts a high level view of the software architecture of the prototype SourceFinder as it currently exists. It has several components – a text gatherer, a text extractor, a text selector and indexer, and a user interface. The text gatherer is essentially a webcrawler that automatically retrieves web pages (i.e. candidate source material) from the Internet by starting at a user specified site and branching to linked pages and sites. The text extractor takes the pages, which are in a variety of formats (e.g. html, pdf), converts them to plain text (ASCII), and segments each page into candidate source paragraphs. The text selector then extracts a variety of linguistic features from each candidate paragraph, including word-frequency, lexical syntactic, and lexical semantic information and uses this information to assign a probabilistic suitability score to each paragraph. SourceFinder currently uses approximately 50 measures to rate text that it finds. The main measures include (a) a set of statistical characteristics (e.g. characteristic ranges of number of words per sentence and number of sentences per paragraph), (b) lexical analyses including a simple measure of argumentation density based upon characteristic argumentation words and phrases, and (c) a readability measure (the Dale-Chall measure; McCallum & Peterson, 1982). Sources containing only low-scoring paragraphs are discarded. Those sources that contain highly rated paragraphs are saved in a database for later review.

**Applications**

SourceFinder has potential applications to many aspects of course preparation. For example, constructivist pedagogy emphasizes situating learning activities within rich, real-world, open-ended contexts. Similarly, there has been a call for more authentic assessments that include material closer to problem solving rather than more closed-form multiple-choice questions (Wiggins, 1989). However constructing such complex questions or activities entails crafting suitable problem contexts.

These rich, real-world contexts present challenges for course preparation. Instructors must strike a balance between the situation not seeming contrived yet not so open ended that students cannot begin to complete the activity. At the same time, to keep material fresh, the problem contexts should be drawn from different domains. However, identifying such material can be difficult when the domain is outside one’s own area of expertise. By using material found through linguistically targeted web search (i.e., by specifying the characteristics of the material that is needed), instructors can assure the appropriate level of complexity and (if the source is appropriate) that the material is a true reflection of the domain (and thus not contrived).

Below we explore some potential applications of SourceFinder to course preparation in different fields.
Exercises and assessments of reading comprehension

Several U.S. states have adopted standards for a variety of literacy skills, including reading comprehension. These standards discuss the importance of students’ ability to comprehend a range of grade-level-appropriate material found outside the classroom. For example, the California Language Arts Content Standards specify that such material might include newspapers, magazines, and online information.

But how are instructors to locate appropriate material for use in instruction or assessments? Not just any newspaper or magazine article is a suitable stimulus for reading comprehension. In addition to reading-level requirements, state standards for student comprehension skills imply that passages must contain certain linguistic characteristics in order that student be able to, for example, identify structural patterns such as cause-and-effect. It is these types of linguistic features that SourceFinder uses when identifying appropriate material.

As an approximation of the difficulty faced by instructors in locating appropriate material, one can look to the people who write reading comprehension questions for standardized tests, such as the Graduate Record Examination® (GRE®), Test of English as a Foreign Language® (TOEFL®), and Graduate Management Admission Test® (GMAT®). On these examinations, tests of reading comprehension typically include a stimulus passage followed by several multiple-choice questions. Passages must satisfy a variety of criteria including style and structure criteria, content criteria, and copyright constraints. Currently, sources are selected manually from (paper) books and long articles. It takes substantial time to select likely source material and read it to locate candidate passages. Test writers estimate that as much as 50% of the professional staff time spent on reading comprehension passages is devoted just to locating an appropriate source.

If professionals who are well practiced in the identification of source material spend so much time searching, imagine the difficulties faced by literacy instructors. SourceFinder can help reduce considerably the time needed to locate material suitable for instructional exercises and assessments of reading comprehension.

Initial evaluations of SourceFinder by test writing professionals suggest that users achieve substantial timesavings for locating suitable passages of text. These users needed to locate passages based on such linguistic factors as passage length, argumentation density, reading level as well as category of content. In their first hour of using SourceFinder, users found as many suitable passages as they would typically find over a three-day period using more traditional approaches.

Supplemental readings and assignments in History

Consider the situation in which a history instructor wishes to assign reading material to the class to supplement the material found in the textbook. While a web search might locate material on the appropriate topic, the instructor must manually review the search results to identify material that is accessible by his or her students (i.e., at the appropriate reading level). Other uses for SourceFinder in history instruction include assignments on critical analysis of several historical documents, which might additionally require that the documents contain a degree of argumentation (one of the linguistic features detected by SourceFinder).

For example, let us say that a high school history instructor is beginning a unit on Christopher Columbus. The instructor wishes to assign supplemental reading on the explorer, plus intends that material to serve as a starting point for a research paper assignment. Desirable linguistic properties of the material include a high school reading level and passages that contain discussions of causes and effects (grist for the research paper).

To test SourceFinder’s capacity to locate material of this sort, we conducted a standard web search (using Google™) on “Christopher Columbus” and immediately located the Millersville University website, containing more than 1000 articles on Columbus and the Age of Discovery. For this example, we limited SourceFinder to search for 15 minutes. In that time, SourceFinder analyzed 135 documents, identifying 29 that contained material having the appropriate linguistic properties. SourceFinder stored these 29 documents in a database, providing a means to easily review the length and general content of each selected document (Figure 2). Double-clicking on a document excerpt opens a window containing that document. Paragraphs within the document that meet the specific linguistic criteria are highlighted in red, allowing the instructor to identify the portions of the document that might make particularly good prompts for the assignment.
Problem sets in Statistics

Creating problems for introductory statistics courses presents challenges because of the need for realistic, “meaty” stories that reflect actual situations in a manner understandable to students (i.e., the storyline cannot require too much knowledge of the nonstatistical subject matter). There are many sources of instructional material and complex statistics problems on the web. For one such site, in seven minutes, SourceFinder identified two statistics problems from a set of 66. Such a result would allow an instructor to quickly determine that this source of statistics material is not linguistically at the right level for his/her class, which might lead the instructor to find a different, more appropriate source of materials.

Conclusions

SourceFinder’s role in course preparation is to help instructors more easily identify on-line materials for class assignments. Based on linguistic analyses, the system screens out material that is clearly not of use to the instructor. Thus, the key benefit of SourceFinder is as a filter: the software will eliminate a large amount of material that would have been inappropriate for the intended use, helping instructors to cope with the potentially huge amount of material returned from traditional, content-based web searches. Of course, it is still the providence of the instructor to select among the remaining (smaller) set of material for the particular educational use envisioned.

Beyond the applications for course preparation, the prevalence of text material on the web suggests other educational applications, such as helping instructors or schools decide among competing textbooks. Several textbook publishers provide sections of candidate textbooks on the web. An instructor might use SourceFinder to determine the textbook having the highest proportion of text that matches the criteria of interest to the instructor, such as reading level and density of information.

Linguistically targeted web search holds much promise for alleviating some of the tediousness of course preparation, freeing instructors for the more intellectually challenging aspects of their work. Our initial success with SourceFinder suggests the viability and generality of the approach.
Acknowledgements

The authors thank Bhasha, Inc. for their work on the development of SourceFinder and Krishna Jha for providing the software architecture diagram used in this paper.

References


