ED 425 727 IR 019 215

AUTHOR Scime, Anthony
TITLE Undergraduate Data Mining on the World Wide Web.
PUB DATE 1998-00-00
NOTE 9p.; In: Association of Small Computer Users in Education: Proceedings of the ASCUE Summer Conference (31st, North Myrtle Beach, SC, June 7-11, 1998); see IR 019 201.
PUB TYPE Reports - Descriptive (141) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Assignments; Course Content; Course Objectives; Database Design; Databases; Evaluation Methods; Higher Education; *Information Retrieval; Information Seeking; *Information Skills; Keywords; Relevance (Information Retrieval); Research Papers (Students); *Research Skills; Search Strategies; *Student Research; Undergraduate Study; User Needs (Information); *World Wide Web
IDENTIFIERS Data Warehousing; Ranking; *Relevance (Evaluation); Search Engines; Web Sites

ABSTRACT

Currently available World Wide Web search engines determine a site's qualification as a response to a search request by matching keywords in the request to keywords representing the site. The returned sites are given a score and ranked according to the match on keywords. Many of these retrieved sites can be irrelevant to the user's true information needs. Undergraduate students with information retrieval and computer literacy skills should be able to search the Web to find and extract information relevant to a domain. A course in the fundamentals of information gathering from distributed heterogeneous sites can improve these skills. The course described in this paper examines methods, theories, and techniques and provides practice in information retrieval, categorization, and knowledge discovery from text and other unstructured data sources such as the Internet and World Wide Web. This paper reports on the results of the course and the experiences gained. Highlights include: course objectives; assignments; the research paper search architecture; outlining the research paper; Web data mining; Web site evaluation, including syntactic and semantic methods, as well as computing overall relevance using a repertory grid; design of a "data warehouse," i.e., a subject-specific relational database of highly ranked documents; knowledge discovery from the data warehouse; and writing the research paper. (Author/AEF)

Undergraduate Data Mining on the World Wide Web

Anthony Scime'
Business and Economics Department
Wilson College
1015 Philadelphia Avenue
Chambersburg, PA 17201
(717) 264-4141
ascime@wilson.edu

Abstract

Currently available World Wide Web search engines determine a site's qualification as a response to a search request by matching keywords in the request to keywords representing the site. The returned sites are given a score and ranked according to the match on keywords. Many of these retrieved sites can be irrelevant to the user's true information needs.

Undergraduate students with information retrieval and computer literacy skills should be able to search the Web to find and extract information relevant to a domain. A course in the fundamentals of information gathering from distributed heterogeneous sites can improve these skills. This course examines methods, theories, and techniques and provides practice in information retrieval, categorization, and knowledge discovery from text and other unstructured data sources such as the Internet and the World Wide Web. This paper is a report on the results of such a course and the experiences gained.

Introduction

In this information age, people are inundated with vast amounts of information. Technology has made this inundation possible and placed information sifting and evaluation on the individual. This is most evident on the World Wide Web. Students need to learn the fundamentals of information gathering from the distributed heterogeneous sites that make up the Web. As they go out into the world, they will use the Internet and the World Wide Web to find, assess, and disseminate information [FULT98]. They need to understand theories and techniques for information retrieval, storage and categorization from textual, unstructured data sources. This paper discusses a course that investigates theories and practices skills in finding information and assembling knowledge from the Web.

Course Requirements

A special topics course in basic distributed information retrieval was developed to allow students to learn how to appraise the quality and reduce the quantity of information retrieved from the Web. The course studies the internet, distributed databases, text based database technology, information retrieval, data warehousing, data mining, web page construction using HTML, data validation, and source citation. The objective is to enable the student to use the World Wide Web as a research tool and gain an understanding of the Web's structure. Using Thuraisingham [THUR97] for the distributed data management theory and Caswell [CASW97] for the practical aspects of the Web, the student:

Demonstrates an understanding of the structure of distributed data sources,
Performs basic information retrieval from the Web,
Develops a data evaluation method,
Creates a data warehouse of retrieved Web site addresses and summaries,
Practices knowledge creation through data mining of the data warehouse, and
Places the knowledge back into text form.

As the major assignment, each student selects a topic to research on the World Wide Web. They design and construct a data warehouse type database to store the results of the information search. Based on the sites found and the use of data mining techniques, an 8-10 page research report is written as an HTML document. A second report of 5-7 pages is written describing the information retrieval, evaluation, storage, and data mining techniques used for creation of the first paper. In the process of meeting the course's research paper requirement, the student must develop the concept of the paper, conduct Web searches, evaluate, store, and review the results, and assess the results for new knowledge. The architecture for this process is provided in Figure 1.

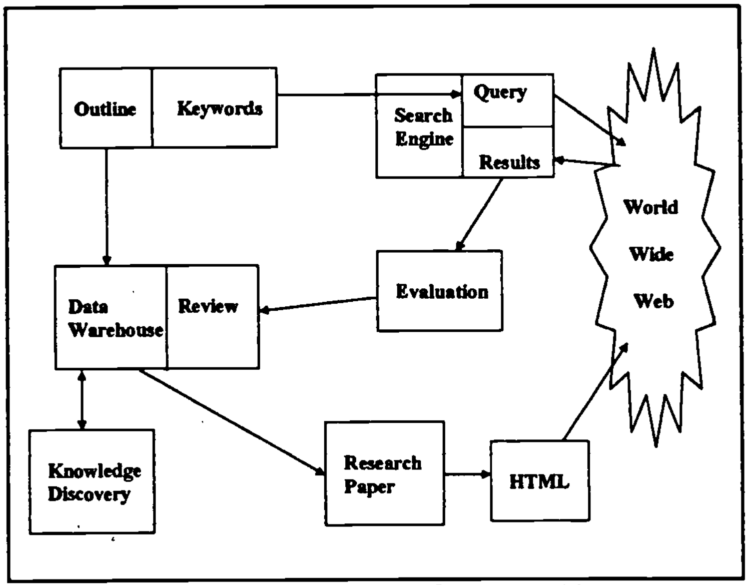

Figure 1. Research Paper Search Architecture

The Data Warehouse is implemented in MS Works 3.0 database or MS Access (for more advanced students). Because a database course is not a prerequisite, the database design consists of a data dictionary, sample table(s) and notional data. During the course, data models are discussed. Students can choose to implement the data warehouse as multiple tables as discussed in this paper. The final populated database contains 40 – 100 entries derived from at least two search engines. Data warehouse design, search methods, evaluation techniques, and data mining methods are determined by the student.

Research Paper Outline and Data Mining the Web

One of the problems with conducting a search is the need to know something about the subject before beginning to research information. The student's initial ideas are expressed as the paper's outline. On the Web, research comes in the form of search engine queries using keywords. The queries should be formulated so that all possible information about the subject will be found. The paper's outline helps define the queries.

Outlining involves the decomposition of a topic into its component parts, which are further decomposed until non-decomposable leaf-node sub-topics are reached. The final sub-topics are defined in an unambiguous, precise manner that avoids confusion as to the intent of the sub-topic. This preciseness will allow for accurate measurement of a document's applicability to the sub-topic. See for example, the outline for an Air Pollution paper in Figure 2.

Topic – Air Pollution	Keywords
1. Environmental Pollution	"Environmental Pollution" or "Air Pollution"
2. Sources	
2.1 Transportation	
2.1.1 Aircraft	"Air Pollution" and Aircraft
2.1.2 Automobiles	
2.1.2.1 Emissions	(Automobiles or Cars) and Emissions
2.1.3 Railroads	"Air Pollution" and Railroads
2.2 Industrial	
2.2.1 Oil Refineries	"Air Pollution" and Industrial and "Oil Refineries"
2.2.2 Power Generation	"Air Pollution" and "Power Generation"
2.3 Fuels	
2.3.1 Coal	Coal
2.3.2 Gasoline	Gasoline
2.3.3 Nuclear Power	"Nuclear Power"
3. Effects	
3.1 Health	"Air Pollution" and Effects and Health
3.2 Agriculture	"Air Pollution" and Effects and Agriculture
4. Control	
4.1 Industrial	"Air Pollution" and Control and Industrial
4.2 Transportation	"Air Pollution" and Control and Transportation

Figure 2. Air Pollution Paper Outline

Data mining is one of the more significant parts of the course. Data mining occurs in two phases. First, the Web is mined to find sites of interest concerning the subject domain. This mining operation is accomplished using the paper's outline and search engines. A query is created for each of the leaf-node sub-topics in the paper's outline using keywords, which represent the sub-topic concept. See Figure 2 for the leaf-node sub-topic keywords for the Air Pollution topic. The queries may be a single keyword, a collection of keywords, a string, or a combination of keywords and strings.

Although a leaf-node sub-topic may have a specific meaning in the context of the paper, the use of a keyword or string could lead to the retrieval of many irrelevant sites [BRAY96]. In this case, keywords and strings may be "and-ed" as far back along the branch path as necessary to increase the specificity of the retrievals, as in sub-topics aircraft, oil refineries, and health (Figure

2). "And" between keywords requires that both keywords occur in the result. For example, the Health sub-topic search using the WebCrawler search engine and the keywords: "Air Pollution" and Effects and Health returns more specific results (310 documents) than simply Health alone (62,757 documents). Because no search engine completely indexes the Web, it is best to use more than one search engine [SEL95]. After evaluation, data on selected sites found are added to the data warehouse.

Evaluation

As sites are found by the search engines, some method is needed to quickly evaluate the site's quality. This effort could be completed programmatically on all the sites returned by the search engine. However, these students have no programming experience. Moreover, manual evaluation forces the student to "think" like the computer. She must evaluate the sites independently of knowledge in the domain. She must use the techniques she developed for evaluation as if she was the computer. Manual evaluation does limit the number of sites that can be considered. Time is a factor in completing the paper. Thus, the order provided by the search engine influences the evaluation results.

Each student develops her own evaluation method. Among possible methods are those based on the syntax of the site and the semantics of the research problem. The first, syntactic, uses heuristics to determine the quality of the site's information. The second, semantic, considers knowledge of the domain from which the search keyword is derived. These techniques are intended for use prior to reading the document. One or a combination of techniques can be used to select document data to populate the data warehouse. Below are provided two sample evaluation methods.

Syntactic Evaluation of a Site

Considering just the document and making a quality evaluation based on its internal structure constitutes a syntactic evaluation. A simple syntactic technique is to assign a value based the URL domain type. The domain type is the last two or three characters of the URL. The most common domain types are ".com" for a business, ".edu" for a school, ".gov" a government agency, ".mil" the military, ".net" a network provider, and ".org" an organization. Two letter domain types refer to the country of the Web site, such as ".us" and ".ca" (United States and Canada).

Checking the three-character domain type may provide a hint about the quality of the information contained at a site. The types ".gov" and ".mil" may be trusted to present government policy and positions on issues.

Information in domain types business and organization may provide information bias to the owner's goals. Business sites are often marketing tools and present the business' products in the best possible manner. Organizations generally have an agenda that is supported on their Web sites.

Schools can present unbiased information on a subject in their search for truth and learning. Quality at ".edu" sites can still be a problem requiring additional syntactic methods. There may be a difference between a site containing a university researcher's published work and a progressive elementary school's Web posting of student papers. Perhaps this can be accomplished syntactically by looking for clues such as "Dr", or "PhD" in the author portion of the document.

A scoring system may be developed where a 3 is assigned to ".edu" PhD sites, 2 to ".gov", ".org" and other ".edu" sites, 1 to ".com", and 0 to any others. With the 10 selections scoring highest having their documents read by the student and their data added to the data warehouse.

Applying the syntactic rules and scoring above on the first 2 of 310 health sub-topic results, the site EPA's Indoor Air Quality Home Page with URL

http://www.epa.gov/iaq/

scores 2 because it is a ".gov" site. Air Pollution and Health with URL

http://www.tec.org/greenbeat/may96/health.html

is an ".org" site and also scores 2.

Semantic Evaluation of Sites

Semantic evaluation considers the value of the site found within the conceptual intent of the query. Search engines use semantics in their query processing by the match of keywords.

A simple semantic method is to count the occurrences of the search keywords in the document title. The more keywords present in the title the higher the score. The 10 sites with the highest scores are entered into the data warehouse.

The top scoring sites are added to the data warehouse. A semantic evaluation on the top two sites from the Health sub-topic gives the first site, EPA's Indoor Air Quality Home Page, a 0 score. None of the keywords "Air Pollution", Effects, or Health, appears in the title. The Air Pollution and Health site scores 2, two of the keywords are present.

Repertory Grid

An improvement in evaluation occurs if the syntactic and semantic scores are combined to provide a site's overall relevance to a user's search request. The product of the syntactic and semantic scores can be computed using a repertory grid [SHA87]. The repertory grid score for the EPA's Indoor Air Quality Home Page, 0, is the product of the syntactic score, 2, and the semantic score, 0. The second site found by the search engine, Air Pollution and Health has a repertory grid score of 4 (2 x 2). This result places the search engine's second place site before its first place site.

Performing this analysis on all sites returned by each search creates rankings that consider the returned documents quality and content. The documents scoring highest for each sub-topic search are added to the data warehouse.

Data Warehouse

A data warehouse is a storage place for data awaiting use. It is a subject specific relational database, which has been populated from diverse and possibly heterogeneous data sources such as exist on the Web. In the operation of the data warehouse, data is selected for inclusion but never sent back to the original source. Transfer is in one direction only. The selection of data items for inclusion is dependent on the purpose for the data warehouse [ROB97]. The previous evaluation phase accomplishes this selection.

As source sites are found and evaluated, a record is written containing the sub-topic of the paper's outline and a quote or summary from the source or the student's ideas on the sub-topic as discussed in the source. The summary record includes additional keywords provided by the student. These additional keywords are generated by the student as she reviews the site. The summary keywords may be different than the keywords used in the site's selection by the search engine. These summary keywords provide additional specificity on the information in the site, as it applies to the overall topic of the research paper. These are details that may not have been evident without the focused reading of the student.

This summary record also contains a bibliographic code to link this record to a second record of bibliographic data. The bibliographic record contains bibliographic information about the source site. This bibliographic record permits the retrieval of the source document should it be needed again. As well, the collection of bibliographic records provides data for the bibliography at the end of the paper. The bibliographic record contains the source's author, title, publisher, copyright date, URL, retrieval date, search engine used for retrieval, the keywords or phases on which the retrieval was based, and the bibliographic code (used to connect to the summary card).

In the example, the Air Pollution and Health document is eligible for entry into the data warehouse. The Air Pollution and Health document found for the sub-topic Health is summarized and results in summary keywords: lungs, ozone, smog, Texas, cars, emissions, and "reduction of air pollution." The keyword and summary (below) becomes part of the summary record in the data warehouse.

> Air Pollution causes a lack of oxygen being taken into the lungs. The air containing large amounts of ozone, sulfur dioxide, nitrogen dioxide, carbon monoxide, lead, and particulate matter reduces the amount of room for oxygen. There is a high correlation between lung disease and air pollution. The less capacity the lungs have, the greater the effect of pollution, such as in children and the elderly. Urban areas in Texas have been found to have especially high levels of ozone pollution. Smog caused by ozone and chlorofluorocarbons is pollution that can be seen by the naked eye. High levels of smog have caused death in some cases. Automobiles are the greatest cause of smog and air pollution. By controlling automobile emissions and increasing fuel efficiency, air pollution can be controlled. Other controls are to reduce the production of electricity by coal.

Knowledge Discovery – Data Mining the Data Warehouse

The second phase of data mining is knowledge discovery from the data warehouse. Data mining is the process of extracting information from large quantities of data. Queries are designed to explore relationships between data records that are not necessarily obvious. Data mining, hopefully, results in nuggets of knowledge heretofore unknown.

This is analogous to gold mining. The prospector through experience and expertise selects a location likely to produce gold nuggets. She proceeds to dig, sift, and pan until gold is found or she realizes this is not the place for a gold mine.

The research paper data warehouse is mined to discover connections between the documents, and their application to the paper outline. This may result in a possible reorganization of the paper. This data mining attempts to classify, associate, or cluster the sites [YEVI97] (Figure 3).

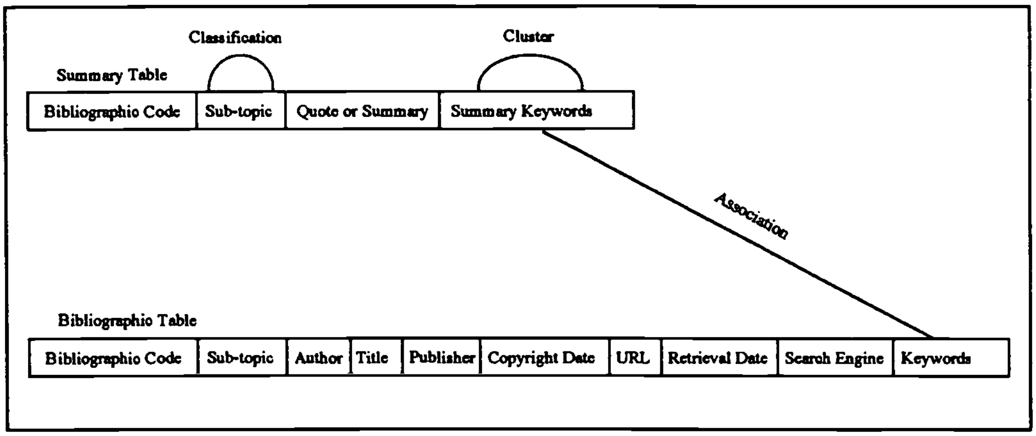

Figure 3. Data Mining the Research Paper Data Warehouse

Classification is the assignment of a database record to a class defined by the database user. For the research paper data warehouse, the paper's structure, its outline, provides the classes of interest. The retrieval keywords used, coming from the paper outline, created groups of documents based on the sub-topics of the outline.

Associations are found by discovering a common value in two data items in different record types. When the keyword in a bibliographic record matches a summary keyword in a summary record, a second document becomes a possible source of information for the original document's sub-topic without regard to the bibliographic code relationship. An association table can be added to the data warehouse.

In clustering, a summary keyword of the same value occurs in records in different classes, different sub-topic values. Documents in the same cluster (using the same summary keyword) are identified by the addition of a cluster table to the data warehouse. Of course, a document may belong to multiple clusters.

For example, data mining the air pollution data warehouse discovers the document "Air Pollution and Health" also fits under the Emissions sub-topic. There is an association between the "Air Pollution and Health" document and the Emissions sub-topic. The "Air Pollution and Health" document contains the summary keywords cars and emissions. These summary keywords are the keywords used in the Emissions sub-topic search. Data mining of the data warehouse discovered this otherwise unknown relationship.

Writing the Research Paper

After complete analysis, old and new structures for the paper appear. The old structure is the outline and the classifications originally developed by the student. The new structure includes associations, and clusters as a result of data mining the data warehouse. The paper is now ready to be written.

The actual writing process is a review of the summaries of the documents in accordance with the paper's outline while being aware of the associations and clusters. The summaries were written by the student after the evaluation process, creating the summary keywords. The summary keywords are the representations of the student's understanding of the concepts in the document. These summary keywords are the basis for the data mining and paper reorganization. The ideas of all the documents are now organized to permit easy writing of the research paper. In addition, the student has the bibliography completed in the data warehouse. Finally, by writing the paper in HTML and placing it on the Web, all the source documents can be reached through hypertext links in the paper.

References

[BRAY96] Bray, T., and Weinberger, D.; "The Hits Just Keep on Comin'"; http://www.opentext.com/corp/otm_hits.html; 1996.

[CASW97] Caswell, S.; New Perspectives on The Internet Using Netscape Navigator Software: Introductory; Course Technology; Cambridge; 1997.

[FULT98] Fulton, K.; "Learning in a Digital Age: Insights intro the Issues;" T.H.E. Journal; 25, 7; February 1998; pp. 60-63.

[ROB97] Rob, P. and Coronel, C.; Database Systems: Design, Implementation, and Management; Course Technology; Boston; 1997 pp. 677-746.

[SEL95] Selberg, E., and Etzioni, O.; "Multi-Service Search and Comparison Using the MetaCrawler"; Proceedings of the 1995 World Wide Web Conference; 1995.

[SHA87] Shaw, M. L. G., and Gaines, B. R.; "KITTEN: Knowledge Initiation and Transfer Tools for Experts and Novices;" Int. J. Man-Machine Studies; September 1987.

[THUR97] Thuraisingham, B.; Data Management Systems; CRC Press; New York; 1997.

[YEVI97] Yevich, R.; "Data Mining;" in Data Warehouse: Practical Advice from the Experts, ed. by Bischoff, J. and Alexander, T.; Prentice Hall; Upper Saddle River, NJ; 1997; pp. 309-321.