Efthimiadis, E. N. and Hendry, D. G. 2005. Search engines and how students think they work. In *Proceedings of the 28th Annual international ACM SIGIR Conference on Research and Development in information Retrieval* (Salvador, Brazil, August 15 - 19, 2005). SIGIR '05. ACM, New York, NY, 595-596. DOI= http://doi.acm.org/10.1145/1076034.1076145

Search Engines and How Students Think They Work

Efthimis N. Efthimiadis and David G. Hendry
The Information School
University of Washington
Seattle, WA 98195, USA
+1 (206) 616-6077 and 616-2316
{efthimis, dhendry}@u.washington.edu

ABSTRACT

To investigate the nature of people's understandings for how search engines work, we collected data from 232 undergraduate and graduate students. Students were asked to "draw a labeled sketch of how search engines work." A reference model was constructed and each sketch was analyzed and compared against it for completeness. The paper presents preliminary results and discusses the implications for educational assessment and curriculum design on the one hand, and information system design on the other.

Categories and Subject Descriptors

H.1 MODELS AND PRINCIPLES: H.1.2 User/Machine Systems; H.3.3 Information Search and Retrieval; H.3.5 Online Information Services, *Web-based services;* K.3.2 Computer and Information Science Education

General Terms

Design, Human Factors.

Keywords

World Wide Web and hypermedia, database access, information retrieval, user and cognitive models, search engines, user studies.

1. INTRODUCTION

In February 2004, approximately 279 million people living in the USA visited Google, MSN, and Yahoo! at least once and performed a search from home, work, or school [5]. Meanwhile, over 50 million people in the USA published content to the web, which is potentially available to the search engines [4]. By these measures alone, search engines, and Google in particular, have become an important cultural phenomenon for their mediation between the everyday producers and consumers of information.

The networked infrastructure that enables information services like Google is an *artificial world* [6]. Like the natural world, it consists of elementary building blocks and intricate structures of enormous diversity. To list just a few: Web pages, keywords, meta tags, hyperlinks, caches, web servers, robots.txt, file permissions, search engines, spiders, users, content providers, advertisers. While human-made, this is not a neat world. Like the natural world, we can normally engage this artificial world without understanding or even being aware of its underlying complexity. Yet, when something does not work as expected we must appeal to its underlying workings. For example, why does my webpage not appear on the first page of Google when I type my name? To answer this question one needs to draw on existing technical knowledge and established concepts and principles. Such knowledge is acquired through interaction with the search engines, coursework, or readings of academic or popular literature.

The question we address in this article is the nature of this technical knowledge held by students of information science. We assert that knowledge of basic technical concepts for search engines is an important kind of literacy. Certainly, this technical knowledge informs how people inquire into noteworthy phenomena, how people assist others with questions about search engines, and how people advocate the use of search engines. For educators, as we are, it is therefore important to take measure of students' knowledge for search engines so that we can design better instruction and be more effective teachers.

2. METHODOLOGY

2.1 Task

Undergraduate and graduate students at the University of Washington were prompted to draw sketches on 8 x 11 in. paper of how search engines work. Students were given approximately 10 min to complete the task at the start of regularly scheduled classes. Participation was voluntary and anonymous. A sample of 232 sketches was collected for analysis.

2.2 Participants

The student participants (N = 232) were assigned to the following three groups: Undergraduate-freshman (n = 53), Undergraduate-informatics (n = 95); and Graduate-information-science (n = 84). While these categories represent three general levels of academic achievement, the demographic profiles for the participants within these groups are heterogeneous, especially for the second two categories, with broad ranges in ages, work experiences, and educational achievement.

2.3 Model of Internet Search

To analyze the sketches, a conceptual model for search was created. The aim of this model, which draws upon standard textbook models of search engines (e.g., [1]), is to identify the major conceptual components of search. The model divides search into three phases with a total of fourteen processing components.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

Copyright is held by the author/owner(s).
SIGIR'05, August 15–19, 2005, Salvador, Brazil.
ACM 1-59593-034-5/05/0008.

A. INDEXING: Processing documents so they can be retrieved later. Components: *Content*; *Spidering/Crawling*; *Parsing*: *Inverted index creation; Link-analysis; Storage.*

B. SEARCHING: Users formulate a query and inspect results
Components: *User; User-need; Query; Results.*

C. MATCHING: Queries are matched against web pages
Components: *Query processing*; *Matching*; *Accessing inverted file*; *Ranking.*

This model is used to assess the completeness of the participants' conceptual models. Raters coded all 232 sketches. The procedural details and results of this analysis will be reported elsewhere [3].

3. RESULTS

3.1 Sketches

The full sample of sketches reveals a tremendous diversity of approaches for explaining the operation of search engines. Some sketches were representational, other systems-oriented, and yet other used computational processes.

3.2 Concept Analysis

The normative model was used to assess the overall presence of the concepts in the sketches. The process for coding the sketches followed these steps. The normative model was documented and discussed by a group of four coders. Working independently, the coders coded a sample of four sketches by inspecting each sketch and voting for the presence or absence of each of the 14 concepts. For sufficient consistency the coders met 3 times to review each others' votes and discuss any differences in judgment. Working independently, each of the 4 coders inspected each of the 232 sketches for the 14 concepts. This resulted in 12,992 votes for the presence or absence of concepts.

The votes were analyzed for inter-coder reliability by computing the percentage of agreed votes between each coder for each concept in each sketch (M = 0.84, SD = 0.02). One or more judges voted differently on the presence or absence of a concept in approximately 16% of the 3,248 concepts considered. Cohen's kappa averaged for all 4 coders is 0.57, which is considered fair [2]. To address this unreliability, the following cut-offs were established: 1) Concept present, if 3 or 4 votes; 2) Concept absent, if 0 or 1 votes; and 3) Concept uncertain, if 2 votes. Then, the authors examined all those identified as concept uncertain and reconciled the disagreements. Using these cut-offs, the votes were counted to determine the presence-or-absence status of each concept in each sketch. This transformed data is used in the analysis below.

3.3 Distribution of Concepts

Figure 1 shows the distribution of the 14 concepts in the 232 sketches. The mean number of concepts per sketch is 4.5 (SD 2.3) with a median of 4.0. The range of concepts covered per sketch is 0 to 13. About 15.5% of the sketches had 4 concepts, 14.2% had 5 concepts, 11.6% had 3 concepts, and 10.8% had no concepts. The mean number of concepts per group ranged from 1 for the freshman, 4 for the Informatics undergraduates and 6 for the graduate students.

3.4 Presence of Concepts

The distribution of the presence of the 14 concepts in the sketches is shown in Figure 2. From this we can clearly see that some concepts, like query, results, content, and match, are more prevalent in the sketches. These are simple concepts that all users understand from the most basic interaction with search systems. However, when we move to the more esoteric concepts, like inverted files, query parsing, link analysis, then we see that very few sketches include them.

Figure 1: Distribution of Concepts

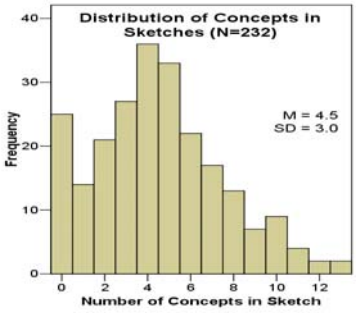
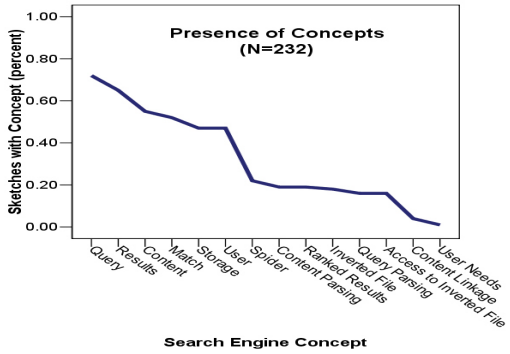


Figure 2: Presence of concepts



4. CONCLUSION

This study presents preliminary results on how people search engines work. The results have both educational and system design implications. In future work we would investigate the development of an educational assessment tool that would provide a reliable method to assess student knowledge on information retrieval concepts and suggest areas for curriculum modification. Further we would develop teaching modules that would help in a) clarifying or correcting misconceptions, and b) helping build richer conceptual models. With respect to search system design issues, a number of questions arise that need further investigation: Does a correct model lead to improvements in: a) search, b) informed consumers, c) becoming sophisticated users of information systems, d) providing a better "feel of control" of the search process? Answers to these questions will help with the redesign of the user interface to facilitate better interaction and better user experience.

5. REFERENCES

[1] Belew, R.K. *Finding out about.* Cambridge University Press, New York, 2000.

[2] Fleiss, J.L. *Statistical Methods for Rates and Proportions*. John Wiley and Sons, New York, NY, 1981.

[3] Hendry, D.G. and Efthimiadis, E.N. Conceptual models for search engines. *Submitted forpublication.*

[4] Lenhart, A., *et al*. Content Creation Online, Pew Internet & American Life Project, 2004.

[5] Media Metrix, c. Top 50 U.S. Internet Property Rankings for Feb 2004. Press Release, Reston, Va. 2004.

[6] Simon, H. *The Sciences of the Artificial.* MIT Press, Cambridge, MA, 1996.