Exhibit 5

Implementation of the SMART Information Retrieval System[*]

Chris Buckley

TR 85-686
May 1985

Department of Computer Science
Cornell University
Ithaca. New York 14853

[*] This study was supported in part by the National Science Foundation under grant IST 83-16166.

Implementation of the SMART Information Retrieval System

Chris Buckley
Department of Computer Science
Cornell University
Ithaca, New York 14853

This study was supported in part by the National Science Foundation under grant IST 83-16166.

1. Introduction

The SMART information retrieval package is a set of programs composing a fully automatic document retrieval system. It allows easy creation, maintenance, and use of on-line document collections. As more information is being kept on-line every day; it becomes more essential to have methods of easy, natural access to the information. The SMART package is primarily a tool for investigating some of these methods. In addition, it is quite usable itself for many applications.

The current SMART system is a collection of programs written in 'C' under the UNIX operating system. It is operational at Cornell University on at least three machines: DEC VAX 11-750, DEC VAX 11-780, and SMI SUN (a 68000 based super micro-computer).

What is described here is only the latest in a long series of SMART implementations; the earliest one being in the early 1960's [1,2,3,4,5]. This, however, is the first implementation allowing actual practical use of the system by a naive user. Previous versions were for experimental purposes only.

This new version naturally draws very heavily on the older versions for its algorithms, although almost no code remains from those versions. A special debt is owed to Ed Fox's implementation which immediately preceded this one [6]. His was the first UNIX implementation, and many of the lessons learned during his work were very useful here.

The current implementation of the SMART system is covered in the remainder of this paper. First, the features and goals of the system are described. The information retrieval process in general is then related to the particular modules and programs within the system. The overall approach to accessing information and parameters is discussed, followed by a brief look at some of the internal data structures used. There is a short section on concurrency, consistency, and protection features. The conclusion discusses the future of the SMART system: what should be done (or re-done) and what could be done.

2. Goals, Features, and Requirements

This implementation of SMART contains few new or radical concepts. Instead, it attempts to provide a solid framework for future work in information retrieval. The two major goals of the current version are to

1. Provide a flexible experimental system for research in information retrieval. See [6] for a discussion of desirable system capabilities and design principles for experimental work.

2. Provide a fast, portable, interactive environment for actual users.

These two goals naturally conflict with each other; the current SMART design is an attempt to satisfy each as much as possible.

The system is concerned with three major types of users: the experimenters, the database administrators, and the naive users. The experimenters need the ability to easily change system parameters and to easily add or replace program modules. The database administrators must be able to create and maintain a collection of documents without worrying about the peculiarities of the particular

2. Providing a help facility for UNIX. There was a lot of documentation for UNIX on-line that was inaccessible because nobody could find it.

3. Accessing a user information database (interests and hobbies as well as factual information).

4. Accessing reference databases (easy, non-factual searches of standard databases of references)

5. Searching electronic mail files (eg. the old mail to system support staff)

6. Searching archives of electronic bulletin boards (USENET news)

The major changes in SMART for the next few years, though, will probably come from the addition of new methods of retrieval, information storage, and models of information retrieval. As experimental work is done, new algorithms will be implemented and added to the present core of the SMART package. At Cornell there are already a number of programs which could augment SMART (eg. clustering, probabilistic retrieval, phrasing). After they are "fine-tuned" a bit more, they will undoubtedly be added. There are still entire areas of information retrieval not covered by the current system. There is now hope for great improvements in the understanding of natural language in information retrieval contexts. Hopefully, the current system can serve as a stepping stone for further research for a number of years to come.

References

[1] J. Rocchio, "Possible Time-Sharing Orgaization for a SMART Retrieval System". Information Storage and Retrieval, Vol 7, Cornell University (1964).

[2] M. E. Lesk, "Design Considerations for Time Shared Automatic Documentation Centers". Information Storage and Retrieval, Vol 11, Cornell University (1966)

[2] M. E. Lesk, "Design Considerations for Time Shared Automatic Documentation Centers". Information Storage and Retrieval, Vol 11, Cornell University (1966)

[3] D. Williamson, R. Williamson, "A Prototype On-line Document Retrieval System" Information Storage and Retrieval, Vol 18, Cornell University (1970)

[4] D. Williamson, R. Williamson, M. E. Lesk, "The Cornell Implementation of the SMART System". In *The SMART Retrieval System*. edited by G. Salton. Prentice-Hall, Englewood Cliffs, N.J. (1971).

[5] SMART Staff, "User's Manual for the SMART Information Retrieval System". Technical Report 71-95, Revised April 1974. Cornell University (1974).

[6] E. Fox, "Some Considerations for Implementing the SMART Information Retrieval System under UNIX". Technical Report 83-560, Cornell University (1983).

[7] G. Salton, M. McGill, *Introduction to Modern Information Retrieval.* McGraw-Hill, New York. (1983).