Characterization of Two New Experimental Collections in Computer and Information Science Containing Textual and Bibliographic Concepts

Edward A. Fox*

83-561
September 1983

*Department of Computer Science
Cornell University
Ithaca, New York 14853

now at:
Department of Computer Science
Virginia Tech
Blacksburg, VA 24061

This work was supported in part by the National Science Foundation, under grant IST-81-08696.

TABLE OF CONTENTS

1 Introduction 2

1.1 Extended Vectors 3

1.2 Contrast with Other Collections 4

1.2.1 Document Collections 5

1.2.2 Query Collections 6

2 Multiple Concept Types 0

3 CACM 13

3.1 Documents 4

3.2 Illustrative Charts and Figures 16

3.3 Queries 0

3.4 Relevance Judgments 41

3.5 Retrieval Performance 4

4 ISI 44

4.1 Documents 4

4.1.1 Background 44

4.1.2 Tape Information Provided 5

4.1.3 Collection Preparation 47

4.2 Subvectors 9

4.3 Queries 56

4.4 Relevance Judgments 8

4.5 Retrieval Performance 58

5 Conclusion 9

References 61

CHARACTERIZATION OF TWO NEW EXPERIMENTAL COLLECTIONS IN COMPUTER AND INFORMATION SCIENCE CONTAINING TEXTUAL AND BIBLIOGRAPHIC CONCEPTS

Edward A. Fox*

Abstract

Two new collections are described which are particularly useful for investigating the interaction between textual and bibliographic data in the automatic indexing and retrieval of documents. An extension to the vector space model has been proposed whereby various types of concepts are included in the representation of such documents. Experiments using an enhanced version of the SMART system have shown such an extended model to perform better than simpler schemes. The CACM and ISI collections developed for this research should be of value for future related studies.

The ISI collection has author, title/abstract, and co-citation data for the 1460 most higly cited articles and manuscripts in information science in the 1969-1977 period. The CACM collection contains 7 types of concepts for the 3204 articles published in the *Communications of the ACM* up through 1979. These collections have 76 and 52 queries, respectively, along with relevance judgments.

*Department of Computer Science, Cornell University, Ithaca, NY 14853; now at Dept. of Computer Science, Virginia Tech, Blacksburg, VA 24061. This work was supported in part by the National Science Foundation, under grant IST-81-08696.

1. Introduction

In order to retrieve documents relevant to the request of a particular user it is necessary to first index or represent the content of articles and manuscripts. For many years this has been done by trained indexers who assign keyword lists or sets of descriptors from a controlled vocabulary [Borko & Bernier 1978]. Since the early 1960's an alternative method of automatic indexing has been developed whereby word stems, words, phrases, or thesaurus category indicators are selected from the title and abstract and a weighted vector indicating the importance of each is constructed [Salton 1980]. Part of this report deals with the vectors derived in this fashion from two collections in information and computer science.

Another source of data about documents is from their bibliographic references. Citation indexes can be used to locate those entries referred to by an article, or which cite it [Garfield 1964, 1979]. Linkages between documents based on bibliographic coupling [Kessler 1962] and co-citation counts [Small 1973] have been utilized for a variety of analysis and retrieval purposes (e.g., [Bichteler & Eaton 1980], [Garfield 1970], [Kessler 1963a, 1963b, 1965], [Small & Koenig 1977], [Small 1978, 1980, 1981], [Weinberg 1974]. Preliminary experimentation has shown that the vectors produced by automatic indexing of document texts can be usefully supplemented by bibliographic information to produce a representation that can be more effectively searched than if either component were used alone ([Michelson et al. 1971], [Salton 1963, 1971]).

To facilitate exploration of the effects of extending the vector space model to include a variety of types of concepts it was necessary to have test collections containing such concepts. One collection containing 1460 of the most highly cited documents in information science published between 1969 and 1977 [Small 1981] was developed based on citation and co-citation data provided by the Institute for Scientific Information. This ISI collection contains three types of concepts: author names, word stems

DOCKET ALARM

Explore Litigation Insights

Docket Alarm provides insights to develop a more informed litigation strategy and the peace of mind of knowing you're on top of things.

Real-Time Litigation Alerts

Keep your litigation team up-to-date with **real-time alerts** and advanced team management tools built for the enterprise, all while greatly reducing PACER spend.

Our comprehensive service means we can handle Federal, State, and Administrative courts across the country.

Advanced Docket Research

With over 230 million records, Docket Alarm's cloud-native docket research platform finds what other services can't. Coverage includes Federal, State, plus PTAB, TTAB, ITC and NLRB decisions, all in one place.

Identify arguments that have been successful in the past with full text, pinpoint searching. Link to case law cited within any court document via Fastcase.

Analytics At Your Fingertips

Learn what happened the last time a particular judge, opposing counsel or company faced cases similar to yours.

Advanced out-of-the-box PTAB and TTAB analytics are always at your fingertips.

API

Docket Alarm offers a powerful API (application programming interface) to developers that want to integrate case filings into their apps.

LAW FIRMS

Build custom dashboards for your attorneys and clients with live data direct from the court.

Automate many repetitive legal tasks like conflict checks, document management, and marketing.

FINANCIAL INSTITUTIONS

Litigation and bankruptcy checks for companies and debtors.

E-DISCOVERY AND LEGAL VENDORS

Sync your system to PACER to automate legal marketing.

WHAT WILL YOU BUILD? | **sales@docketalarm.com** | **1-866-77-FASTCASE**

fastcase®
Smarter legal research.