An Interactive Iterative Method for Electronic Searching of Large Literature Databases

by
Marco A. Hernandez

Submitted in partial fulfillment
of the requirements for the degree of
Doctor of Professional Studies
in Computing

at

School of Computer Science and Information Systems

Pace University

April 2013

We hereby certify that this dissertation, submitted by Marco A. Hernandez satisfies the dissertation requirements for the degree of *Doctor of Professional Studies in Computing* and has been approved.

_____-_____

Fred Grossman, Ph.D. 12 April 2013
Chairperson of Dissertation Committee

_____-_____

Chuck Tappert, Ph.D. 12 April 2013
Dissertation Committee Member

_____-_____

Lixing Tao, Ph.D. 12 April 2013
Dissertation Committee Member

School of Computer Science and Information Systems
Pace University 2013

Abstract

An Interactive Iterative Method for Electronic Searching of Large Literature Databases

by
Marco A. Hernandez

Submitted in partial fulfillment
of the requirements for the degree of
Doctor of Professional Studies
in Computing

April 2013

PubMed® is an on-line literature database hosted by the U.S. National Library of Medicine. Containing over 21 million citations for biomedical literature--both abstracts and full text--in the areas of the life sciences, behavioral studies, chemistry, and bioengineering, PubMed® represents an important tool for researchers.

PubMed® searches return a list of citations based on keywords. Given the amount of information available through PubMed, this study asks, "How do you leverage your search to look for subtle relationships between documents?" Data suggest that users rely on only the first page or returned results. This presents a non-trivial problem.

One reason is there is no formal standard or syntax, which has been identified to denote these relationships; another is that the nature of these relationships is ambiguous. Computational search models mimic techniques from Library Sciences, which anticipate that users usually do not fully understand, or cannot completely articulate their needs. Lacking complete information, librarians approximate need and return relevant documents in an iterative fashion.

Current search engines rely on mathematical modeling for text retrieval. At a fundamental level, this involves the use of a dictionary or ontology for term refinement, word stemming (i.e. removing *ing*), and clustering analysis to represent the relationship between terms and documents. The nuances of the interaction with a librarian are lost in this transaction.

This study created a tool, Iterative Matrix Search (IMS) that uses biomedical ontologies and electronic lexicons in order to include closely related words in a search. The tool returns an expanded key word list and a list of the documents where hits were found, as well as a normalized score outlining the relationship between the documents and key words.

The study then surveyed life sciences researchers to gain an understanding of how they search for information. They submitted unsuccessful queries previously done against PubMed®.

IMS expanded the participants' initial queries and provided them with formatted results that allowed for connections through word associations.

A follow-up interview determined researchers' perceptions of tool's utility in cataloging new information, knowledge, or insight.

The technique demonstrated benefits of IMS to enable novel insights and relationships in the literature.

Acknowledgements

Christine D'Onofrio, my loving spouse, and Amelia Hernandez, my daughter, for their patience and support through this sometimes dark dissertation process.

Mario and Victoria Hernandez, my parents, who taught me the value of an education

Table of Contents

List of Tables

List of Figures

Chapter 1 Introduction

The convergence of telecommunications, information sciences, and computer science and engineering, has facilitated an explosion in the quality and quantity of information that is available to individuals on any number of different media. The dissemination of this data occurs as fast as our communications infrastructure can propagate it. We can now carry the collective knowledge of our species in a pocket or purse (i.e. smartphones and wireless data networks).

The sheer volume of information available through the Internet and its rate of growth make our ability to glean insight or knowledge, find patterns in what (on the surface) is unrelated data, and make connections that lead to hypothesis validation or creation, challenging at best.

For the first time in human history, the problem we face is information abundance, thus shifting the focus from accessing information to filtering, collating, and interpreting that information. A typical WWW or archive search returns tens of thousands of documents. This amount of information is impractical and time consuming for a user to sort through to determine which document is the most relevant to the query, especially when the summarization provided on a results page may not properly represent the content, or, as often happens, the question is ambiguous.

This study is about Information Retrieval (IR), which is about finding information. It encompasses *information need*, *search*, *retrieval* and *access*. Specifically:

> *Information retrieval is the process of matching the query against the information objects that are indexed. An index is an optimized data structure that is built on top of the information objects, allowing faster access for the search process. The indexer tokenizes the text (parsing), removes words with little semantic value (so-called stop-words), and unifies word families (so-called stemming). The same is done for the query as well. Users express their information need as a request (search terms) and it is formulized as a query for the retrieval system. The information system responds by matching information objects, which are relevant to this query. Information retrieval focuses on finding relevant information rather than simple pattern matching. It is also important to note that relevance is a subjective notion, since different users may make various judgments about the relevance or non-relevance of particular documents or information objects to given questions.*[33]

Biomedical research in the 21st century is largely a data driven discipline. Current experimental protocols can produce terabyte levels of output. The biomedical literature consists of scientific papers reporting new results integrated with experimental data and husbanded through the peer-review process. The biomedical literature validates the researcher's hypothesis and opens the door to expanding knowledge. The techniques detailed in this study leverage the entirety of the returned results in a query, rather than simply the first page, and allow users to re-order the list interactively, based upon their assessment of returned results.