

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

UMI Number: 3569888

All rights reserved

INFORMATION TO ALL USERS

The quality of this reproduction is dependent upon the quality of the copy submitted.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if material had to be removed, a note will indicate the deletion.



UMI 3569888

Published by ProQuest LLC (2013). Copyright in the Dissertation held by the Author.

Microform Edition © ProQuest LLC.

All rights reserved. This work is protected against unauthorized copying under Title 17, United States Code



ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 - 1346

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.
Chairperson of Dissertation Committee

12 April 2013

Chuck Tappert, Ph.D.
Dissertation Committee Member

12 April 2013

Lixing Tao, Ph.D.
Dissertation Committee Member

12 April 2013

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.

PREVIEW

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

PREVIEW

Table of Contents

Abstract	iii
List of Tables	xvii
List of Figures	xxi
Chapter 1 Introduction.....	1
1.1 Search and Knowledge	4
1.2 ‘Typical’ Search engines.....	8
1.2.1 What do Traditional Search Engines Provide?	8
1.2.2 What Content Gets Noticed	9
1.2.2.1 Traditional Search Engine Output Format	11
1.2.2.2 Phase Structure Grammar (i.e. making connections based on linguistic relationships).....	13
1.2.3 Search Engine Overview.....	14
1.2.3.1 Traditional Search Algorithms.....	15
1.2.3.2 Google’s® Search Algorithm	15
1.2.3.3 PubMed’s® Search Algorithm.....	16
1.2.4 Iterative Matrixed Search (IMS) Developed for this Study.....	17
1.2.5 Comparison of IMS and standard search engines	18
1.2.6 Other Search Engines.....	19
1.3 How to Take Data to Knowledge: an Interdisciplinary Problem.....	19
1.3.1 Leveraging Semantic Frames in Search Results	20
1.4 Problem Statement	21
1.4.1 Problem 1. Understanding the Sources of Information and Search Behaviors in Biomedical Research (Search Baseline).	21

1.4.2 Problem 2. Sifting Through Pages of Output, a More Functional Method to Connect with Data.....	22
1.4.2.1 How IMS Works and Meets Research goals	23
1.4.2.2 IMS Step 1 - Expanding the Query	24
1.4.2.3 IMS Step 2 – Performing a Lexical Search with the Expanded Query Terms	25
1.4.2.4 IMS Step 3 – Ranking the Results from I2E.....	26
1.4.2.5 IMS Step 4 – Ordering Results by Relevance based on User Input ..	28
1.4.2.6 IMS Review	30
1.4.3 Problem 3, Effectively Reducing the Size of the Data Indexed Without Losing Information (Noun Phrases).....	31
1.4.3.1 Using the Noun Phrase (NP) as a Lexical Construct to Reduce Computational Overhead	31
1.4.4 Why Biomedical Research and PubMed®	31
1.4.5 The Need to Leverage Unstructured Data	33
1.4.5.1 Issues around Searching Unstructured Data	34
1.5 Study Goals Achieved.....	35
1.5.1.1 Understanding Search Behavior in Biomedical Research	35
1.6 Significance of Research.....	36
1.6.1 Application of the Research Outside of the Biomedical Sciences.....	38
1.7 Research Questions.....	38
1.8 Limitations of the Study.....	39
1.9 Definition of Terms.....	40
Chapter 2 Relevance of Research in the Context of Other Work.....	44
2.1 Historical Overview	44
2.1.1 The World Wide Web and Information Retrieval	47

2.1.2	The Importance of Text Mining in Biomedical Research	48
2.1.3	Institutional Support within the Biomedical Research Community	51
2.1.4	Molecular Biology and Data Mining	51
2.2	Research Literature	53
2.2.1	Document Search	54
2.2.1.1	Document Search Algorithms	55
2.2.1.2	Document Term Weighting Approaches	59
2.2.2	Technologies and Algorithms to Address Document Search	61
2.2.3	Term Frequency – Inverse Document Frequency (TF-IDF)	62
2.2.3.1	TF-IDF in Search Engines	63
2.2.4	Latent Semantic Analysis (LSA) in Information Retrieval	64
2.2.4.1	Overview of Latent Semantic Analysis	65
2.2.4.2	Singular Value Decomposition (SVD)	66
2.2.4.3	Dimensionality Reduction	67
2.2.4.4	Query Projection and Matching in LSA	68
2.2.4.5	Weighting Algorithms in LSA	68
2.2.5	Vector Space Distances	69
2.2.6	Latent Semantic Analysis (LSA): Practical Use	72
2.2.7	Manipulating the Result Vectors	78
2.2.8	Latent Semantic Analysis Summary	79
2.3	How People Search	80
2.3.1	Search Tactics	80
2.3.2	The QIRO Model Scales Up	82
2.3.3	Search Patterns	83
2.3.4	Interacting and Interfacing with Search	84

2.4	Lexicons and Ontologies.....	85
2.4.1	Electronic Lexicons	86
2.4.1.1	WordNet®.....	87
2.4.1.2	WordNet® vs. Traditional Dictionary Organization	89
2.4.1.3	The WordNet® Lexical Matrix.....	93
2.4.1.4	Representing meaning in WordNet®.....	95
2.4.1.5	Synonymy	96
2.4.1.6	Antonymy	97
2.4.1.7	Hyponymy.....	97
2.4.1.8	Meronymy.....	98
2.4.1.9	WordNet® and Linguistic Morphology.....	98
2.4.2	Mathematical Formalization of Semantic/Knowledge Models	101
2.4.2.1	Lattice Theory.....	101
2.4.2.2	Formal Concept Analysis.....	103
2.4.2.3	Summary of Lattice Theory and Formal Concept Analysis	104
2.4.3	Other Electronic Lexicons, an Overview	105
2.4.3.1	VerbNet.....	105
2.4.3.2	PropBank.....	106
2.4.3.3	FrameNet.....	106
2.4.3.4	OntoNotes	112
2.4.3.5	TreeBank.....	113
2.5	Ontologies	114
2.5.1	Ontology Background Information.....	115
2.5.2	How Ontologies Function	117
2.5.3	Ontologies and the WWW	119

2.5.4	How is Knowledge Codified.....	121
2.5.5	Ontological Organization and Knowledge Capture.....	121
2.5.5.1	Upper Ontology	122
2.5.5.2	Middle Ontology.....	125
2.5.5.3	Domain Ontology.....	125
2.5.6	Ontological Organization.....	126
2.5.6.1	Taxonomic Ontologies.....	127
2.5.6.2	Descriptive Ontologies.....	129
2.5.6.3	Space and Time Descriptive Ontologies.....	129
2.5.7	Ontologies vs. Databases	130
2.5.8	Representing Expert Knowledge	130
2.5.8.1	First Order Logic.....	131
2.5.8.2	Description Logic (DL).....	131
2.5.8.3	Open World and Closed World Semantics	132
2.5.8.4	Examples of Description Logic	132
2.5.8.5	Description Logic Constructs	136
2.5.8.6	Description Logic Summary	137
2.5.9	Horn Logic	137
2.5.10	Frame Logic	138
2.5.11	Frame Logic and TBox / ABox	139
2.5.12	Description and Frame Logic Comparisons	140
2.5.13	Web Ontology (AKA the Semantic Web)	143
2.5.13.1	Resource Description Format (RDF)	146
2.5.13.2	SPARQL Protocol and RDF Query Language	148
2.5.13.3	Web Ontology Language – OWL	149

2.6	The Role of Grammatical Structures in the Study	150
2.6.1	Introduction.....	150
2.6.2	What is a Grammatical Phrase?	151
2.6.3	Noun Phrase	151
2.6.4	Other Linguistic Constructs	152
2.7	PubMed® Corpus	155
2.7.1	Medline	155
2.7.2	PubMed® vs. MEDLINE®	157
2.7.3	MeSH (Medical Subject Heading) and PubMed®	158
2.7.4	PubMed® XML Tag Format	159
2.7.5	PubMed® Search Engine.....	160
2.7.5.1	How PubMed® Works	162
2.8	Contribution this study will make to the field	165
Chapter 3	Study of Search in Biomedical Research Survey Results.....	167
3.1	Research Methods Employed	167
3.1.1	Purpose of the On-Line Questionnaire	168
3.1.2	Purpose of the In-Person Interview.....	168
3.1.3	Research Design and Methods.....	169
3.1.4	Development of Survey for Biomedical Professionals.....	172
3.1.5	Data Collection Methodology.....	173
3.1.6	Participant Protection.....	174
3.1.7	Participant Selection Procedures.....	175
3.1.8	Data Analysis Procedures	175
3.2	How Biomedical Researchers Search (Electronic Survey Results)	176
3.2.1	Who are the Biomedical Researchers in this study.....	176

3.2.2	What Types of Search Behaviors do Biomedical Researchers Exhibit?	179
3.2.3	What Sources do Biomedical Researchers Use?	183
3.2.4	How Often is Search Used in Biomedical Research.....	184
3.2.5	Offline Sources of Knowledge for Biomedical Researchers	184
3.2.6	Why do Biomedical Researchers use Search.....	185
3.2.7	How do Biomedical Researchers Choose Their Data Sources?	186
3.2.8	How Long Do Biomedical Researchers Spend on Search.....	187
3.2.9	What are the Search Frustrations among Biomedical Researchers	188
3.3	Biomedical Research Search Usage Summary	189
3.3.1	National Library of Medicine PubMed® Log Analysis Summary	189
3.3.2	Summary of Study Survey on Biomedical Research Search	190
3.4	Formats for Presenting Questionnaire Results.....	193
3.5	Tools and Databases Used in the Electronic Survey Questionnaire	193
3.6	Electronic Survey Conclusions.....	194
3.7	In-Person Interview.....	195
3.7.1	Background Information on Biomedical Researchers in the Study	196
3.7.1.1	Interview Background Section Q1. How Long Have You Been at Your Current Position	196
3.7.1.2	Interview Background Section Q2. How Long is Your Professional Experience in Life Sciences.....	197
3.7.1.3	Interview Background Section Q3. Current Job Title	199
3.7.1.4	Interview Background Section Q4. Organizational Role	199
3.7.2	Technical Comfort	200
3.7.2.1	Interview Technology Comfort Section Q5. General Knowledge and Comfort with technology.	200
3.7.2.2	Interview Technology Comfort Section Q6. Microsoft Software Comfort Level.....	201

3.7.2.3	Interview Technology Comfort Section Q7. Job-Related Time Spent on the Internet	201
3.7.2.4	Interview Technology Comfort Section Q8. Frequently Visited WWW Sites	202
3.7.2.5	Interview Technology Comfort Section Q9 and Q10. Favorite and Least Favorite WWW Sites.	203
3.7.2.6	Interview Technology Comfort Section Q11 & Q12. Physical Environment.....	204
3.7.2.7	Interview Technology Comfort Section Q13. Work Related Limits on Information Access	205
3.7.3	Work-Related Tasks.....	205
3.7.4	Search Behavior (How Do You Search).....	206
3.7.4.1	Search Behavior Section Q 22. Do You Always Use the Sites Identified in Question 21.....	206
3.7.4.2	Search Behavior Section Q23. Alternative Sites for Information.....	207
3.7.4.3	Search Behavior Section Q24. What You Know About the Topic and What You're Looking for	207
3.7.4.4	Search Behavior Q25. Types of Information Needed.....	208
3.7.4.5	Search Behavior Section Q26. Using Different WWW Sites for Different Needs.....	208
3.7.4.6	Search Behavior Section Q 27 & Q29. Search Strategies and Evaluating Results	208
3.7.4.7	Search Behavior Section Q28. Use of Advance Search Features.....	210
3.7.4.8	Search Behavior Section Q30. How Do You Determine Relevance	212
3.7.4.9	Search Behavior Section Q31. Data Elements Focused on to Determine Relevancy.....	213
3.7.4.10	Search Behavior Q 32. Once You Have Results, What Are the Next Actions taken	214
3.7.5	In-Person Interview Summary	214
3.8	In-person Interview Conclusions	216

Chapter 4	IMS In-Person Tutorial	218
4.1	Introduction.....	218
4.2	Tools and Algorithms Used in IMS	219
4.2.1	Latent Semantic Analysis	219
4.2.2	Using Synonyms to Expand Search Terms.....	220
4.2.3	I2E Lexical Search Engine.....	221
4.2.4	Microsoft Excel®.....	221
4.3	Standardized Query Used for Demonstration and Training	222
4.3.1	Definition of the Terms in the Standardized Query.....	222
4.3.2	Justification of Standardized Query.....	222
4.4	Steps used in IMS Standardized Query	223
4.4.1	Step 1- PubMed® Query	223
4.4.2	Purpose of Running IMS Against the Standardized Query	224
4.4.3	IMS Step 1 – Expanding the Query Terms.....	226
4.4.4	IMS Step 2 – Semantic Search.....	227
4.4.5	IMS Step 3 – Latent Semantic Analysis	230
4.4.6	IMS Step 4 – Interactive IMS Search	235
4.4.7	Re-Ordering the Result List in the Standardized Query	236
4.5	Comparison of Different Search Parameters using I2E	237
4.6	Tools and Database Used in This Study	238
4.7	Literature Analysis Tools Evaluation	241
4.8	Conclusion	241
Chapter 5	IMS Study Query Results	243
5.1	Introduction.....	243
5.1.1	IMS Results Summary:.....	243

5.2	Queries Returning Useful Data.....	244
5.2.1	Research Query 1: Roux-en-Y and its effect on Type II Diabetes (Gastric By-Pass Surgery and Diabetes).....	245
5.2.1.1	Background.....	246
5.2.1.2	Difference with Roux-en-Y	247
5.2.1.3	Research Query 1 - Searches	248
5.2.1.4	Research Query 1 - Results.....	249
5.2.1.5	Research Query 1 - Conclusions.....	258
5.2.2	Research Query 2: What Genes Are Involved in PIK3CA – AKT1 Co-activation.....	260
5.2.2.1	Research Query 2 – Searches.....	261
5.2.2.2	Research Query 2 - Results.....	263
5.2.2.3	Research Query 2 - Conclusions.....	263
5.2.3	Research Query 3: HCV Protease Inhibitor Adverse Event	265
5.2.3.1	Research Query - 3 Searches	268
5.2.3.2	Research Query 3 - Results.....	268
5.2.3.3	Research Query 3 - Conclusions.....	272
5.2.4	Research Query 4: Testicular Toxicity	273
5.2.4.1	Research Query 4 – Searches.....	274
5.2.4.2	Research Query 4 - Conclusion	276
5.2.5	Research Query 5: DPPx Inhibition.....	277
5.2.5.1	Research Query 5 - Searches	277
5.2.5.2	Research Query 5 - Results.....	280
5.2.5.3	Research Query 5 - Conclusions.....	282
5.3	Sample of a Search that did not Return Successful Results.....	282
5.3.1	Research Query 6: Interleukin 3 and 4 Relationship with Asthma.....	283

Chapter 6	Conclusions.....	286
6.1	Study Background.....	286
6.2	Initial Study Goals	287
6.2.1	Research Question: How Important Is Search in Biomedical Research... 287	
6.2.2	How is search conducted (What are the Search Behaviors Exhibited).....	288
6.2.3	Expanding Search Through The Use Of Subject Ontologies And Lexical Databases	289
6.2.4	Allowing Iterative Search and Fast Browsing by Exposing the Query Matrices.....	289
6.3	Conclusions of Study	290
6.3.1	Comparison of Survey Results to PubMed.....	291
6.3.2	Search Behavior	292
6.3.3	Interactive Matrix Search (IMS) Technique	294
6.3.4	Noun Phrases	295
6.4	Implications.....	295
6.5	Recommendations.....	299
6.5.1	Data Input.....	300
6.5.2	Thesauri / Ontology	300
6.5.3	User Interface.....	301
6.6	Recommendation for Further Study.....	301
Glossary	WordNet Glossary of Terms.....	303
Appendix A	On-Line Research Questionnaire	308
Appendix B	In Person Interview.....	316
Appendix C	Participant Recruitment Correspondence	322
Citations	324

List of Tables

Table 1 Google Search Results vs. Click-Through.....	10
Table 2 Search Engine Comparison.....	18
Table 3 I2E Output.....	25
Table 4 Sample Term by Document Matrix	27
Table 5 Word Frequency List	28
Table 6 Ordered Result List.....	29
Table 7 Examples of Biomedical Text Mining.....	52
Table 8 Sample Word-by-Document Matrix	57
Table 9 Sample Query Vector.....	58
Table 10 Local Weighting Techniques	68
Table 11 Global Weighting Techniques	69
Table 12 An Example, Three Documents and Query	74
Table 13 Term-By-Document Matrix and Query Vector	75
Table 14 Rank Order Matrix.....	77
Table 15 WordNet® 3.0 Statistics	93
Table 16 Lexical Matrix.....	95
Table 17 Morphy Suffixes and Endings	99
Table 18 Formal Concept Lattice	102
Table 19 FrameNet Logical Frame Elements	108
Table 20 FrameNet Semantic Roles	111
Table 21 TreeBank Thematic Role Assignments	114
Table 22 Comparison of Frame and Description Logic Systems	141

Table 23 Pre-Determined List of Protein Interactions	154
Table 24 MEDLINE® Statistics	156
Table 25 PubMed DTD.....	160
Table 26 Participant Professional Background	176
Table 27 Participant Age	177
Table 28 Participant Role.....	177
Table 29 Participant Professional Experience	178
Table 30 Participant Search Experience	179
Table 31 Participant Search Behavior.....	181
Table 32 Sources for Electronic Search in order of preference	183
Table 33 Search Frequency.....	184
Table 34 Offline Search Activities	185
Table 35 Search Reasons and Motivation in Ranked Order	185
Table 36 Search Strategies.....	186
Table 37 Factors for Choosing an Information Source.....	187
Table 38 Factors for not choosing an Information Source	187
Table 39 Search Duration (minutes).....	188
Table 40 Search Frustrations	188
Table 41 Tenure at Current Job Position	197
Table 42 Life Sciences Industry Experience	198
Table 43 Current Job Title	199
Table 44 Current Job Roles.....	200
Table 45 Technology Comfort Level.....	201
Table 46 Microsoft Office Comfort Level.....	201
Table 47 Daily Time Spent on the Internet.....	202

Table 48 Frequently Visited WWW Sites	203
Table 49 Location of Online Research	204
Table 50 Physical Office Space	204
Table 51 Limits on Access to Information	205
Table 52 Site Re-Use Frequency	206
Table 53 Alternative Sources of Information	207
Table 54 Results Data Types	207
Table 55 Search Information Types.....	208
Table 56 Initial Search Behavior (Interview)	210
Table 57 Advanced Search Features.....	211
Table 58 Crucial Search Functionality	212
Table 59 Result Relevancy	213
Table 60 Relevant Meta Data Elements.....	213
Table 61 Post Search Actions	214
Table 62 I2E Output.....	229
Table 63 Excerpt of Word by Document Matrix for Standardized Query	231
Table 64 Excerpt of Word Frequency List for Standardized Query	232
Table 65 Excerpted Query Vector	233
Table 66 LSA Ordered Document List.....	234
Table 67 Re-Ordered Result List	236
Table 68 Effects of Term Expansion on I2E	237
Table 69 I2E Query Results for RYGB	250
Table 70 RYGB Word Frequency Vector	252
Table 71 RYGB Term by Document	253
Table 72 Rank Ordered RYGB Matrix.....	255

Table 73 PubMed vs. IMS (Rue en Y Query)	259
Table 74 PubMed vs. IMS PI3K AKT Co-Activation.....	264
Table 75 PubMed vs. IMS HCV Protease Inhibitor Adverse Events	272
Table 76 PubMed vs. IMS Testicular Toxicity Results	277
Table 77 PubMed vs. IMS DPP Inhibition Results	282
Table 78 PubMed vs. IMS DPPx Inhibition	285
Table 79 IMS Results Summary Table	291

PREVIEW

List of Figures

Figure 1 Position of Clicked References (Creative Commons License, NIH)	11
Figure 2 List Based Search Results	12
Figure 3 PubMed based search results on Alzheimer's Disease.....	23
Figure 4 Query and Document Representation (derived from Goker)	70
Figure 5 Term Document Vectors	71
Figure 6 Search Interaction Model	85
Figure 7 Formal Concept Line Diagram.....	103
Figure 8 The Ontology Spectrum	118
Figure 9 Hierarchy of Ontologies	122
Figure 10 The Beer Ontology	126
Figure 11 Evolution of Ontologies	127
Figure 12 Biological Taxonomies of Life.....	128
Figure 13 Open vs. Closed World Models.....	133
Figure 14 Concept Map	135
Figure 15 Inference Diagram.....	136
Figure 16 The Semantic Web Stack (Derived from T. Berners-Lee).....	144
Figure 17 The Semantic Model.....	146
Figure 18 PubMed Standardized Query Result Excerpt.....	224
Figure 19 IMS Workflow.....	226
Figure 20 Roux-en-Y (courtesy of the U.S. National Library of Medicine (Creative Commons))	248
Figure 21 DPP 4 Inhibitions (Derived from National Library of Medicine).....	279
Figure 22 PubMed Search Results.....	298

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.