

Representing Scientific Knowledge

Chaomei Chen · Min Song

Representing Scientific Knowledge

The Role of Uncertainty

Chaomei Chen
College of Computing and Informatics
Drexel University
Philadelphia, PA
USA

Min Song
Department of Library and Information
Science
Yonsei University
Seoul
South Korea

ISBN 978-3-319-62541-6 ISBN 978-3-319-62543-0 (eBook)
<https://doi.org/10.1007/978-3-319-62543-0>

Library of Congress Control Number: 2017957689

© Springer International Publishing AG 2017

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

The 2014 Ebola outbreak in West Africa raised many urgent concerns about public health and safety, as well as legal and administrative implications. The high mortality rate of the Ebola virus heightened the tension between the public, healthcare providers, patients, and local authorities. In the United States, the White House expressed concerns about possible unintended consequences of quarantine policies enforced on doctors and nurses returning from Ebola-stricken countries. Governors of some states defended their quarantine policies, while the White House worried that the policies might not be grounded in science. Some contractors were deeply concerned about the safety of handling Ebola patients' medical wastes and whether they should discard expensive instruments just because they were used to analyze Ebola patients' blood. Some people firmly believed that people without symptoms of Ebola would not transmit the disease. However, the Center for Disease Control and Prevention (CDC) revising its own guidelines was enough reason for others to take extra prudent measures to minimize the risk.

Charles Haas, an environmental engineering professor at Drexel University, specializes in water treatment and risk assessment. He started his comprehensive search in the literature for any information on how long the Ebola virus might be able to survive in water. He did not find a clear answer in the literature. Instead, he found reports of nonzero probabilities of infection after 21 days, which was the basis for the recommended 21-day quarantine. Similarly, a group of researchers did a deep search in the literature but did not find a clear picture either. The implications of these findings on public health policies, public understanding of science, and information science are striking.

Semantic MEDLINE is a great resource for developing a good understanding of scientific knowledge in terms of semantic predications as well as their original unstructured texts. The complexity of scientific writing is strikingly high. It is common to see long and complex sentences. Studying semantic predications and the contexts in which they appear has revealed how frequently uncertainties go hand in hand with the very knowledge one aims to achieve. Knowledge that is free from uncertainty probably has no value in a research field. Understanding the

epistemic status of scientific knowledge is so important that we want to claim that expertise is the knowledge of uncertainty!

The profound and integral role of uncertainty in science, especially in research fronts of a scientific field, has become the core interest of our research. In April and December 2016, two workshops in association with the National Center for Science and Engineering Statistics (NCSES) of the NSF catalyzed the focus on uncertainty further. At the April workshop with the NCSES, Chen presented some of the initial ideas and preliminary results of uncertainties associated with scientific publications in a white paper on the fidelity of visualizing scientific uncertainty.

The preparation and launch of a new open access journal, *Frontiers in Research Metrics and Analytics* (RMA), in midsummer of 2016 provided another boost to the idea. While many have pointed out the shortcomings of overly relying on simplistic and often single metrics of research productivity and quality, evaluators and policymakers are currently limited to only a few options. As a result, it is difficult to compensate the lack of semantic, diagnostic, and analytic reasoning due to over-simplifications of scientific inquiry as a complex adaptive system. The mission of RMA is therefore to bridge the currently loosely coupled research communities. The theme of improving the clarity of the epistemic status of science emerges again in the five grand challenges for accessing and communicating scientific knowledge more efficiently and effectively.

The idea of creating a Visual Analytic Observatory of Scientific Knowledge (VAO) becomes a unifying framework to stimulate and accommodate tools, resources, and applications toward meeting the five grand challenges and beyond. The research project led by Chen is supported by the NSF Science of Science and Innovation Policy (SciSIP) program (Award Number 1633286). The VAO aims to enable researchers to find the epistemic status of scientific knowledge and its provenance of evolution efficiently and effectively. With the worldwide user community of our CiteSpace tool, we believe that the VAO will substantially advance the state of the art. This book introduces the theoretical foundations of how scientific fields develop, which the reader can then use as a referential framework to guide subsequent explorations of scientific knowledge. We also introduce science mapping tools and demonstrate how these tools can help us develop a better understanding of the history and the state of the art of a scientific domain. More importantly, we want to share our methods and principles, both theoretical and practical, with our reader so that we can empower ourselves with computational techniques and analytic reasoning. In particular, creativity comes from competing, contradictory, and controversial views. Reconciliations of existing discrepancies may lead to creative solutions at a higher level. We hope our reader can benefit from the analytic and methodological value of the materials presented in the book.

Chen spent his sabbatical leave at Yonsei University in the Spring semester of 2017 and taught two courses on Yonsei University's beautiful campus on visual exploration of scientific literature. Students from these two classes eagerly and diligently explored and applied the science mapping tools we introduced in this book, namely, CiteSpace, VOSviewer, and CitNetExplorer.

In our previous work, we emphasized the pitfalls and biases of mental models in our reasoning and decision-making. In this book, we aim to demonstrate that uncertainty plays a fundamental role in representing and communicating scientific knowledge.

We are truly grateful for the encouragement and support from many people at various stages of our research and the production of the book. Chen would like to take this opportunity to thank our coauthor Min Song, researchers in his Text and Social Media Mining Lab (TSMM), and students at Yonsei University for collaborative research and the hospitality during Chen's sabbatical in Seoul. Chen is also grateful to Jianguo He and Qing Ping as graduate research assistants at Drexel University, Sergei V. Kalinin at Oak Ridge National Laboratory for exploring applications of science mapping in material sciences and for organizing a tutorial in Boston, Maryann Feldman for encouragements and guidance, Gali Halevi, Henk Moed, and Mike Taylor for their valuable contributions toward research on tracking emerging trends, Caroline S. Wagner for organizing one of the workshops with NCSSES and serving as a guest editor for a Research Topic with RMA, and Jie Li at Shanghai Maritime University for his extensive efforts in disseminating science mapping tools in China. We would like to say thank you to Beverley Ford at Springer for her initiative, encouragement, persistence, and patience.

As always, to the members of Chen's loving family, Baohuan Zhang, Calvin Chen, and Steven Chen, thank you for everything.

Acknowledgements

Chaomei Chen wishes to acknowledge the support of the NSF SciSIP Program (Award Number 1633286) and industrial sponsorship in the past from Elsevier and IMS Health.

Min Song acknowledges the support of the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2015S1A3A2046711).

Philadelphia, USA
Seoul, South Korea
August 2017

Chaomei Chen
Min Song

Endorsement

Chaomei Chen and Min Song have written an important book that opens up a new area in the study in scientometrics and informetrics as well as information visualization, namely the study and measurement of uncertainty of scientific knowledge and how uncertainty is expressed in scientific texts. At the same time the book is a tutorial and review of relevant methods in natural language processing and gives step by step instructions on how they can be implemented. What I like most about the book, however, is how it integrates this new approach with existing theories in the history and sociology of science. In my view, uncertainty is key to understanding the development of scientific knowledge.

Henry Small
Senior Scientist
SciTech Strategies

Contents

1	The Uncertainty of Science: Navigating Through the Unknown	1
	Introduction	1
	Mount Kilimanjaro	2
	Heilmeier’s Catechism	6
	How Does Science Advance?	8
	Paradigm Shift and Gestalt Switch	9
	Competing for Recognition	11
	The Evolution of a Specialty	13
	Searching for the Unknown	15
	Scientific Controversies	16
	Conflicts and Contradictions in Science	16
	The Mass Extinction Debates	17
	Catastrophism	18
	Gradualism	19
	Public Understanding of Science	20
	The Tower of Babel	20
	Communicating with Aliens?	21
	Controversies in the Ebola Crisis	24
	Laypersons Explanations of Conflicting Scientific Claims	26
	Grand Challenges	28
	Accessibility	28
	Clarity on Uncertainty	29
	Connecting Diverse Perspectives	30
	Benchmarks and Gold Standards	30
	Integrating Scholarly Metrics and Analytics	31
	The Organization of the Book	31
	References	34

2	The Dynamics of Scientific Knowledge: Macroscopic Views	37
	Introduction	37
	Mental Models	38
	Easy to Form	38
	Hard to Change	40
	Theories of Scientific Change	41
	Scientific Revolutions	41
	Paradigm Shift	42
	Criticisms	43
	Explanation Coherence	45
	Competition Leads to Scientific Change	46
	Permanent Discovery	48
	Specialization	49
	Fragmentation	49
	An Evolutionary Model	51
	Multiple Perspectives	53
	Summary	54
	References	55
3	Science Mapping Tools and Applications	57
	Keeping Abreast of Scientific Frontiers	57
	Scholarly Publication	58
	Citation-Based Analysis	59
	The Metaphor of a Knowledge Space	61
	CiteSpace: Visualizing and Analyzing a Knowledge Domain	63
	Visual Exploration of Scientific Literature	66
	Data Collection	67
	Configuration of Representation Models	67
	Link Selection	68
	Node Selection	70
	Interactive Visualizations	72
	Structural Variation Analysis	73
	Using MySQL Databases in CiteSpace	75
	VOSViewer and CitNetExplorer	75
	Terrorism Research (1996–2003)	79
	Citation Bursts	79
	Timeline Visualization	81
	Structural Variations	84
	Terrorism Research (1980–2017)	85
	Semantic Structures of Clusters	90
	Concepts in Context	96
	Main Path Analysis	98
	Structural Variations	104
	Science Mapping	106

The Interplay Between Science and Theories of Science	106
Characterizing the Field of Study	107
Visual Analysis of the Literature	108
Visualizing the Field	111
Landscape View	113
Timeline View	115
Major Specialties	116
Cluster #0—Science Mapping	116
Cluster #1—Domain Analysis	121
Cluster #2—Research Evaluation	124
Cluster #3—Information Visualization and Visual Analytics	126
Trajectories of Citations Across Cluster Boundaries	128
Trajectories of Prolific Authors	129
Articles with Transformative Potentials	129
The Emergence of a Specialty	131
Summary	134
References	135
4 Measuring Scholarly Impact	139
Introduction	139
Information Metrics	140
Information Content	140
Year-by-Year Labels of a Cluster	143
Selecting Noun Phrases with LSI	143
Selecting Indexing Terms with LSI	145
Semantic Relatedness	146
Resnik's Semantic Similarity	147
Other Measures of Semantic Similarity	149
Concentration	155
Burstness	155
An Automaton	155
Burst Detection in CiteSpace	156
Log-Likelihood Ratio	162
Likelihood Ratio	162
Characterizing a Cluster	163
Entropy	167
Clumping Properties of Content-Bearing Words	170
Condensation	170
Clumping Versus TF*IDF	173
Importance and Impact	173
Degree Centrality and Eigenvector Centrality	173
Hirsch Index	177
The g-Index	178
Other Measures	179

Normalization of Metrics	181
Distributions of Citation Counts	183
Influential Factors on Citations	186
Improvements of Impact Factors	187
Science Mapping	190
Exploring the Science Mapping Dataset with CiteSpace's Database	191
Major Subject Categories in Science Mapping	192
Citation Distributions	196
Citation Normalization Over Time	200
Summary	202
References	202
5 Representing Biomedical Knowledge	205
Introduction	205
MEDLINE	206
MeSH	206
ULMS	208
SemRep	210
Extracting Semantic Predications	212
The Interactive Mode	212
The Batch Mode	215
Semantic MEDLINE	217
References	221
6 Text Mining with Unstructured Text	223
Natural Language Processing	223
Modeling and Analytic Tools	224
Information Extraction	225
Extracting Entities from Text	225
Extracting Entities from Biomedical Literature	227
Extracting Relations from Text	229
Named Entity Recognition	231
Shallow Parsing	232
Negation	232
Feature Construction from Defined Rules	233
ML-Based Classification	233
Well-Known Relation Extraction Tools	234
Open IE	234
Extracting Semantic Predications with SemRep	235
Topic Modeling	239
Latent Semantic Indexing	240
Latent Dirichlet Allocation (LDA)	242
Semantic Networks and Ontology	245

WordNet	247
BabelNet	251
Deep Learning	254
Word Embeddings	255
Summary	259
References	259
7 Literature-Based Discovery	263
Swanson's Pioneering Work	263
Major Trends of LBD	265
LBD Systems	266
ArrowSmith	266
BITOLA	268
Hypothesis Generator	271
PKD4J: A Scalable and Flexible Engine	274
Design Principle	274
Architecture	275
Preprocessing	276
Entity Extraction	276
Relation Extraction	277
Storing the Results of Extraction	278
Recent Developments and Remaining Challenges	280
References	280
8 Patterns and Trends in Semantic Predications	283
Semantic MEDLINE Database	283
Representing Semantic Predications as a Graph	283
Causality Claims on Ebola	291
Conflicting Claims	293
When Was a Causal Relationship Initially Hypothesized?	293
Measuring the Importance of Semantic Predications	294
Contradictions as a Source of Uncertainty	298
Semantic Predications on Virus Research (1914–2014)	300
Exploring a Semantic Network of Predications in CiteSpace	304
Causal Relations in Virus Research	304
Visual Analysis of Semantic Predications	307
Constructing a Semantic Network	308
Option 1: Top N MEDLINE Articles	309
Structural Variations	318
Option 2: MEDLINE Articles by g-Index	325
Structural Variations	330
Summary	335
References	336

9 Visual Analytic Observatory of Scientific Knowledge	337
Introduction	337
Visual Analytic Observatory of Scientific Knowledge	338
Types of Uncertainties in Scientific Literature	340
Hedging and Speculative Cues	341
Finding Semantically Equivalent Uncertainty Cues	345
Citation Distortion and Provenance of Evidence	347
Retraction	348
Distributions of Uncertainty Cues	351
Contradictory Claims	353
The Reduction of Uncertainty	358
Propositions and Their Epistemic Status	360
Dependency Graphs	364
The Length of Uncertain Statements	366
Summary	371
Concluding Remarks	371
References	372

About the Authors

Chaomei Chen is Professor in the College of Computing and Informatics at Drexel University and Professor in the Department of Library and Information Science at Yonsei University. He is the Editor in Chief of Information Visualization and Chief Specialty Editor of Frontiers in Research Metrics and Analytics. His research interests include mapping scientific frontiers, information visualization, visual analytics, and scientometrics. He has designed and developed the widely used CiteSpace visual analytic tool for analyzing patterns and trends in scientific literature. He is the author of several books such as Mapping Scientific Frontiers (Springer), Turning Points (Springer), and The Fitness of Information (Wiley).

Min Song is Underwood Distinguished Professor at Yonsei University. He has extensive experience in research and teaching in text mining and big data analytics at both undergraduate and graduate levels. Min has a particular interest in literature-based knowledge discovery in biomedical domains and its extensions to a broader context such as the social media. He is also interested in developing open source text mining software in Java, notably creating the PKDE4J system to support entity and relation extraction for public knowledge discovery.

List of Figures

Fig. 1.1	Mountain of scientific publications indexed in the Web of Science (as of 2014)	5
Fig. 1.2	This network of co-cited references reveals rapid changes of research topics in Scientometrics (1980–2015)	10
Fig. 1.3	The geographic distribution of authors who published on Ebola	25
Fig. 3.1	The meta-data of a research article—a 2006 JASIST article on CiteSpace II (Chen 2006). The article is the 2nd of the 10 Google Scholar classic papers in Library and Information Science published in 2006.	60
Fig. 3.2	An overview of terrorism research (1996–2003)	65
Fig. 3.3	Shepard’s analysis definitions	72
Fig. 3.4	The structure of the system before the publication of the ground breaking paper by Watts and Strogatz (1998)	74
Fig. 3.5	The structure of the system after the publication of Watts and Strogatz (1998).	74
Fig. 3.6	An interface with MySQL in CiteSpace	76
Fig. 3.7	A density map visualization in VOSViewer of references cited in the science mapping dataset	77
Fig. 3.8	A direct citation network visualized in CitNetExplorer	78
Fig. 3.9	Drill down and the shortest path between two nodes in CitNetExplorer.	78
Fig. 3.10	Articles with citation bursts in terrorism research (1996–2003)	80
Fig. 3.11	A timeline visualization of the terrorism research (1996–2003)	82
Fig. 3.12	Structural variation by transformative link count	85
Fig. 3.13	A cluster-view visualization of the terrorism research (1980–2017). Node selection by g-index ($k = 10$).	86
Fig. 3.14	A network overlay shows the 1996–2003 network in the context of the 1980–2017 network	87

Fig. 3.15	A timeline visualization of terrorism research (1980–2017)	88
Fig. 3.16	A timeline visualization of citation bursts	91
Fig. 3.17	Cluster # New York City (1997–2009)	92
Fig. 3.18	A concept tree of cluster #0 based on terms extracted from the abstracts of its citing articles, i.e. the research front.	92
Fig. 3.19	A concept tree of Cluster #1 biological weapon	93
Fig. 3.20	A concept tree of Cluster #2 terror attack	93
Fig. 3.21	A concept tree of Cluster #3 transnational terrorism based on terms extracted from abstracts of citing articles to the cluster	94
Fig. 3.22	A concept tree of Cluster #4 islamic state	94
Fig. 3.23	Cluster #7—WTC cough syndrome.	94
Fig. 3.24	A concept tree of cluster #7 world trade center cough syndrome.	95
Fig. 3.25	An unfiltered concept tree of the WTC cough syndrome cluster (#7)	95
Fig. 3.26	The contexts of foreign aid in a concept-in-context tree of cluster #3 transnational terrorism.	96
Fig. 3.27	Sentences that mentioned the term democracies in abstracts of Cluster #3 transnational terrorism	97
Fig. 3.28	Some of the contexts of radicalization in cluster # Islamic state	97
Fig. 3.29	Open the directed citation network in Pajek	99
Fig. 3.30	Retain the largest connected component of the network	99
Fig. 3.31	Identify the largest connected component to retain	100
Fig. 3.32	Extract the largest connected component by selecting cluster 2 to retain	100
Fig. 3.33	Shrink strongly connected components	101
Fig. 3.34	Remove loops from the largest connected component.	101
Fig. 3.35	Compute traversal weights along main paths.	101
Fig. 3.36	Create the main paths by including multiple key routes, e.g. 1–10	102
Fig. 3.37	Use the LAYERS.MCR macro to draw the generated main paths.	102
Fig. 3.38	Main paths derived from the direct citation network based on search path count (SPC)	103
Fig. 3.39	The modularity of the baseline network changes over the years	105
Fig. 3.40	Structural variations measured by the relative entropy of the distributions of betweenness centrality.	105
Fig. 3.41	Topic search queries used for data collection.	108
Fig. 3.42	The distribution of the bibliographic records in Set #14	109
Fig. 3.43	The main user interface of CiteSpace	109
Fig. 3.44	A dual-map overlay of the science mapping literature.	111

Fig. 3.45	A hierarchy of indexing terms derived from Set #14.	112
Fig. 3.46	49 references with citation bursts of at least 5 years	113
Fig. 3.47	A landscape view of the co-citation network, generated by top 100 per slice between 1995 and 2016 (LRF = 3, LBY = 8, and $e = 1.0$).	114
Fig. 3.48	A timeline visualization of the largest clusters of the total of 603 clusters.	115
Fig. 3.49	A hierarchy of key concepts selected from citing articles of Cluster #0 by log-likelihood ratio test	118
Fig. 3.50	High impact members of Cluster #0	119
Fig. 3.51	Top 20 most cited references in the largest cluster	120
Fig. 3.52	Major citing articles to the largest cluster	121
Fig. 3.53	A hierarchy of key concepts in Cluster #1.	122
Fig. 3.54	Key members of Cluster #1.	122
Fig. 3.55	Key members of Cluster #1, sorted by sigma	123
Fig. 3.56	Citing articles to Cluster #1.	123
Fig. 3.57	A hierarchy of key concepts in Cluster #2.	125
Fig. 3.58	High impact members of Cluster #2	125
Fig. 3.59	High impact members of Cluster #2	126
Fig. 3.60	Citing articles of Cluster #2.	126
Fig. 3.61	A hierarchy of key concepts in Cluster #3.	127
Fig. 3.62	High impact members of Cluster #3	127
Fig. 3.63	Key members of Cluster #3.	128
Fig. 3.64	Citing articles of Cluster #3.	128
Fig. 3.65	Novel co-citations made by 8 papers of White HD (left) and by 14 papers of Thelwall M (right).	129
Fig. 3.66	Three examples of articles with high modularity change rates: (1) Waltman (2016), (2) Zupic (2015), and (3) Zhu et al. (2015)	132
Fig. 3.67	Stars indicate articles that are both cited and citing articles. Dashed lines indicate novel co-citation links. Illustrated based on 15 papers of the author's own publications	133
Fig. 3.68	Citation trajectories of Howard White's publications and their own locations	133
Fig. 3.69	Novel links made by a review paper of informetrics (Bar-Ilan 2008)	134
Fig. 4.1	Information content of top 50 most common keywords in 17,731 science mapping articles	141
Fig. 4.2	Generating year-by-year labels of a cluster in CiteSpace.	144
Fig. 4.3	Year-by-year labels of the biological terrorism cluster.	144
Fig. 4.4	WS4J Demo at http://ws4jdemo.appspot.com	150
Fig. 4.5	The local structure of dime, nickel, and their LCS in WordNet and intermediate measures used in semantic similarity algorithms	151

Fig. 4.6	The user can modify the automaton by adjusting a few parameters	157
Fig. 4.7	The distributions of three title terms with the strongest bursts.	159
Fig. 4.8	A cluster view of title terms in terrorism research (1990–2017). Term labels are proportional to the strength of their burst. Labels starting with # are cluster labels, e.g. #0 terrorist attacks.	159
Fig. 4.9	Project Demo 1 in CiteSpace. Cluster labels are selected by LLR	167
Fig. 4.10	Compute statistical associations with log-likelihood ratio tests in CiteSpace	168
Fig. 4.11	Associations between title terms articles published in the Scientometrics	168
Fig. 4.12	The information entropy of terms extracted from articles on terrorism each year.	169
Fig. 4.13	High eigenvector centrality nodes are concentrated in Cluster #1 blast over-pressure. Zooming into #1 at the next level reveals high betweenness centrality nodes such as Mallonee1996 and Burns1993.	176
Fig. 4.14	The size of a node represents its betweenness centrality (left), PageRank (middle), and eigenvector centrality (right) . . .	177
Fig. 4.15	A dual-map overlay visualization of the terrorism2017 dataset (N = 14,656 articles and reviews)	180
Fig. 4.16	The main field-level citation paths include Psychology Education Health to Psychology Education Social (z = 8.423, f = 17,276), Economics Economic Political → Economics Economic Political (z = 7.075, f = 14,602)	180
Fig. 4.17	Some of the most cited journals: 1. <i>Journal of Conflict Resolution</i> , 2. <i>Journal of Personality and Social Psychology</i> , 3. <i>Journal of Traumatic Stress</i> , 4. <i>Terrorism and Political Violence</i> , and 5. <i>The American Journal of Psychiatry</i>	181
Fig. 4.18	Distributions of citations by year of publication in <i>Scientometrics</i> (2010–2014).	181
Fig. 4.19	Distributions of the average number of references per paper in Terrorism (1982–2017)	182
Fig. 4.20	A lognormal distribution	185
Fig. 4.21	The number of records in the dataset of Science Mapping (1980–2017)	191
Fig. 4.22	A log-log plot of the frequencies of citations per paper in Science Mapping (1980–2017)	197
Fig. 4.23	Trends of the number of references and the number of citations of the four largest subject categories	199
Fig. 4.24	Cumulative relative citations by year of publication	201

Fig. 4.25	Probabilities of articles published in each year having citations c in Science Mapping	201
Fig. 4.26	5-year moving average of citation probabilities	202
Fig. 5.1	The hierarchical structure of MeSH descriptors	207
Fig. 5.2	The search page of MeSH	208
Fig. 5.3	The result page of the query “Raynaud disease.”	208
Fig. 5.4	An example of Metathesaurus concept	209
Fig. 5.5	Biologic function hierarchy	209
Fig. 5.6	An illustrative example of the UMLS	210
Fig. 5.7	System architecture of SemRep	211
Fig. 5.8	The user interface of the interactive SemRep.	213
Fig. 5.9	Interactive SemRep’s interface.	213
Fig. 5.10	The results from SemRep in the full fielded output format	214
Fig. 5.11	The input file format for SemRep’s batch mode	216
Fig. 5.12	The user interface of the batch mode SemRep.	216
Fig. 5.13	Scheduler batch job status	217
Fig. 5.14	Semantic MEDLINE’s search page	218
Fig. 5.15	Semantic MEDLINE’s result page.	218
Fig. 5.16	A radial layout visualization of a network based on 1000 predications on HIV and AIDS	219
Fig. 5.17	The network contains the CAUSES relations only (lines in red)	220
Fig. 6.1	An example of entity extraction results for location, organization, and person	227
Fig. 6.2	An architecture of a typical supervised relation extraction system	232
Fig. 6.3	A resulting RDF graph for drug reposition	238
Fig. 6.4	Concept map using natural language relationships.	247
Fig. 6.5	The homepage of WordNet search.	250
Fig. 6.6	t-SNE visualizations of word embeddings	256
Fig. 6.7	Similar words associated with the word “IBM” over time.	257
Fig. 6.8	The shift of meanings of words in HistWords.	258
Fig. 7.1	Swanson’s UPK model—the connection of fish oils and Raynaud disease	264
Fig. 7.2	An example of Brat visualization of entity and relation.	266
Fig. 7.3	The homepage of ArrowSmith.	267
Fig. 7.4	The resulting B-term list for “Raynaud disease” and “Fish oil”	268
Fig. 7.5	Filtered B-terms.	269
Fig. 7.6	The search results for the query <i>magnesium</i>	270
Fig. 7.7	The results of the related concepts Z to “Magnesium”	271
Fig. 7.8	The list of related concepts Y to the target term “Magnesium”.	272

Fig. 7.9	The search homepage of the hypothesis generator.	272
Fig. 7.10	The search result page for the query “Raynaud disease”.	273
Fig. 7.11	The results of extracted entities (left) and the path analysis start page (right)	273
Fig. 7.12	The path analysis result.	273
Fig. 7.13	The overall architecture of PKDE4J.	275
Fig. 7.14	Entity extraction component. An extended version of Song et al. (2015).	276
Fig. 7.15	Relation extraction component.	277
Fig. 8.1	A network of semantic predications visualized in Neo4j.	284
Fig. 8.2	A graphical answer to the question: who has published what paper containing sentences that cited which references?	286
Fig. 8.3	A sub-graph containing sentences in which CiteSpace is the subject	287
Fig. 8.4	The earliest causality claims involving Ebola	292
Fig. 8.5	Detecting bursts in semantic predications on causal relations in research on virus	297
Fig. 8.6	Some of the earliest semantic predications found with bursts.	297
Fig. 8.7	Major semantic relations in the conflict versus no conflict sets of articles from Semantic Medline	300
Fig. 8.8	The distribution of semantic predications over time. The red line is the 5-year moving average	301
Fig. 8.9	A total of 38,256 MEDLINE records are converted to a data file for subsequent analysis with CiteSpace	306
Fig. 8.10	A semantic network of 338 UMLS concepts connected by 1158 semantic predications of causality relations (1980–2016). CiteSpace: Top 50. Largest CC: 331 (92%). Q: 0.4125. S: 0.267	310
Fig. 8.11	A timeline view of the semantic predications on causality relations.	311
Fig. 8.12	A close-up to the timeline view of the four largest clusters of predications	312
Fig. 8.13	Most frequently appeared concepts in the virus dataset.	313
Fig. 8.14	The concept hiv and its neighbouring concepts connected through causal connections	314
Fig. 8.15	The HIV concept has a burst of 6.7642 between 1990 and 1992. It appeared in 1190 PubMed records	314
Fig. 8.16	The burstiness of the concept Virus (Strength: 110.9355, duration 1980–1988). The concept appeared in 3481 PubMed records.	315
Fig. 8.17	Explore the source of a semantic predication.	315

Fig. 8.18	A Pathfinder network of predications. 31 clusters labelled by LLR on predications. Node selection: g-index; Link retention: Pathfinder on time-sliced networks and the merged network	316
Fig. 8.19	The concept HIV and neighboring concepts. For example, HIV causes AIDS (0.15)	317
Fig. 8.20	A timeline view of the Pathfinder network. Nodes are selected by their g-index scores	317
Fig. 8.21	Zooming in	318
Fig. 8.22	Novel connections in dashed lines are made by a 1983 article (PubMed ID: 6870184)	319
Fig. 8.23	Trajectories of novel links added by top 10 articles with the strongest modularity change rates	320
Fig. 8.24	A Pathfinder network of semantic predications generated in CiteSpace. Time slicing: 3; TopN: 100; Range: 1990–2014.	321
Fig. 8.25	Structural variation analysis of the semantic predications (1990–2014) in CiteSpace (3-year intervals)	322
Fig. 8.26	A closer view of novel semantic links between distinct clusters made by a MEDLINE article (PMID: 14766405)	323
Fig. 8.27	The semantic predications extracted from the article (PMID: 14766405)	324
Fig. 8.28	The burstness of the concept Epstein-Barr virus between 1992 and 1995	325
Fig. 8.29	A network of co-occurring semantic predications extracted from MEDLINE articles on virus research over 101 years (1914–2014). Node selection was based on the g-index ($k = 10$). Clusters of semantic predications are labelled by semantic predications with all the citing articles	326
Fig. 8.30	A timeline visualization of the semantic predications on causal relations	326
Fig. 8.31	A slightly different view of the timeline visualization with an emphasis on the distribution of predications over time	327
Fig. 8.32	Semantic predications with a period of burst for 25 years or longer	328
Fig. 8.33	The ontological tree of semantic predications in the largest cluster (#0)	328
Fig. 8.34	The distributions of predication 7581872 in the collection of predications on virus (left) and within cluster #0 (right)	329
Fig. 8.35	Distributions of the predication 5292122 within cluster #0 (left) and in the entire virus dataset (right)	329
Fig. 8.36	A hierarchy of major semantic predications in Cluster #4 on relations between hepatitis c virus and the liver disease	330

Fig. 8.37	Distributions of the leading predication in Cluster #4 (left: the entire virus dataset and right: Cluster #4)	330
Fig. 8.38	Structural Variation Analysis: the five dashed lines are novel links introduced by a 1999 MEDLINE article (PMID: 9989205)	331
Fig. 8.39	FEIGAL1999 made transformative links across different clusters	333
Fig. 8.40	A 1972 MEDLINE article (PMID: 4340152) added two novel predications and reinforced one existing predication. The modularity change rate induced by the article is 7.83. It also shifted the distribution of betweenness centrality scores of the nodes by a degree of 0.05	333
Fig. 8.41	The footprint of a 2001 article (PMID: 1134302), which has the largest number of incremental links	334
Fig. 8.42	Incremental links made by article PMID: 1134302 are all within Cluster #0	334
Fig. 9.1	Architecture of a visual analytic observatory of scientific knowledge	339
Fig. 9.2	Hedging words in the conflicting set versus the non-conflicting set	343
Fig. 9.3	Uncertainty cue words from the original seed list (in red) and expanded (in green)	347
Fig. 9.4	Uncertainty cue words colored in 11 semantically equivalent classes. The size of a label is proportional to the eigenvector centrality of the corresponding word	347
Fig. 9.5	Retracted articles (red dots) in a co-citation network.	350
Fig. 9.6	A retracted article by Nakao et al.	350
Fig. 9.7	The first page of the retracted article by Nakao et al. (2003) along with semantic predications extracted from the article	351
Fig. 9.8	Contradicting semantic predications extracted from MEDLINE records on causal relations between HIV and AIDS	356
Fig. 9.9	The diversity of published claims decreases over time.	360
Fig. 9.10	The first appearance of the predication “HIV CAUSES AIDS” in 1984 (PMID: 6095415; SID: 35618164 [1]: “A transmissible agent especially a retrovirus (HTLV, LAV), is now widely considered in AIDS etiology.”).	364
Fig. 9.11	The dependency graph of a sentence in a 1984 article (PMID: 6100647; SID: 30287966 [4]). HTLV-I is etiologically associated with adult T-cell leukemia-lymphoma (ATLL), HTLV-II has been isolated from a patient with hairy T-cell leukemia, and HTLV-III is the cause of acquired immune deficiency syndrome (AIDS)	365

Fig. 9.12	The dependency graph of a sentence from a 1984 article (PMID: 6145881; SID: 35528335[6]). This is the 6th sentence in the abstract: <i>The results strongly indicate that the antibodies to HTLV-III are diagnostic of AIDS or indicate significant risk of the disease, and suggest that HTLV-III is the primary cause of human AIDS.</i>	365
Fig. 9.13	The dependency graph of a sentence from a 1984 article (PMID: 6200936; SID: 34893490[7]). This is the 7th sentence in the abstract: <i>These results and those reported elsewhere in this issue suggest that HTLV-III may be the primary cause of AIDS</i>	366
Fig. 9.14	The dependency graph of a sentence in a 1990 article (PMID: 2104787; SID: 18493183[16]). This is the 16th sentence in the abstract: <i>The results of this study clearly indicate that PBM from HIV+ individuals are endowed with the capacity to mediate ADCC against HIV-infected/coated cells and thus, we postulate that PBM may play a direct role in vivo in lysis or suppression of HIV-coated/infected cells and in the pathogenesis of AIDS</i>	367
Fig. 9.15	A dependency graph of a sentence in a 2009 article (PMID: 19202348; SID: 120435934[12]). This is the 12th sentence in the abstract: <i>The conference opening was memorable for a number of reasons: among these was the presence of South Africa's new Minister of Health, Barbara Hogan who, in her first speech in a major forum as a senior member of the SA Government, affirmed that HIV causes AIDS, and that the search for a vaccine is of paramount importance to SA and the rest of the world</i>	367
Fig. 9.16	The dependency graph of the title of a 1988 article (PMID: 3399880; SID: 20897139 [title]). The title is: <i>HIV is not the cause of AIDS</i>	368
Fig. 9.17	The dependency graph of a sentence of a 1990 article (PMID: 1980675; SID: 51884237[8]). This is the 8th sentence in the abstract: <i>Duesberg recently published that HIV and AIDS may well be correlated, but stated that HIV is not the cause of AIDS</i>	368
Fig. 9.18	The dependency graph of a sentence of a 1996 article (PMID: 8906995; SID: 40872383[8]). This is the 8th sentence in the abstract: <i>Furthermore, Cys-138 was found in chimpanzee immunodeficiency virus (CIV), a lentivirus that is similar to HIV but does not cause AIDS in chimpanzees.</i>	369

- Fig. 9.19 The dependency graph of a sentence from a 2008 article (PMID: 18624032; SID: 111111060[1]). This is the opening sentence of the abstract: *More than a decade ago, the pathogenesis of AIDS was reviewed in this journal, using the subtitle 'classical and alternative views', when evidence was accumulating that HIV could not cause AIDS simply through direct cytopathic mechanisms alone.* 370
- Fig. 9.20 A streamgraph visualization of semantic predications between 1984 and 1989. 370

List of Tables

Table 2.1	Four categories of scientific change.....	48
Table 3.1	Major citing articles of Cluster #0.....	84
Table 3.2	Major citing articles of Cluster #1.....	84
Table 3.3	Major citing articles of Cluster #2.....	84
Table 3.4	Major citing articles of Cluster #0 in Terrorism Research (1980–2017).....	89
Table 3.5	Major citing articles of Cluster #1 in Terrorism Research (1980–2017).....	89
Table 3.6	Major citing articles of Cluster #2 in Terrorism Research (1980–2017).....	90
Table 3.7	Major citing articles of Cluster #3 in Terrorism Research (1980–2017).....	90
Table 3.8	Major citing articles of Cluster #4 in Terrorism Research (1980–2017).....	90
Table 3.9	The five largest clusters of co-cited references of the network of 3145 references.....	116
Table 3.10	Temporal properties of major clusters.....	117
Table 3.11	Potentially transformative papers published in recent years (2012–2016).....	130
Table 4.1	Information content of the most common keywords in a set of science mapping articles.....	142
Table 4.2	Information contents of low-frequency keywords.....	142
Table 4.3	Year-by-year label terms of the biological terrorism cluster.....	145
Table 4.4	Year-by-year cluster labels extracted from indexing terms of the biological terrorism cluster.....	146
Table 4.5	Semantic similarity algorithms with sim (dime, nickel) as an illustrative example.....	152
Table 4.6	Comparing algorithms with Miller and Charles’ (1991) experiment and Resnik’s 1995 experiment.....	153

Table 4.7	The burst durations of 48 title terms between 1990 and 2017 in terrorism research.	157
Table 4.8	The number of articles selected by the g-index each year to construct the network of title terms	161
Table 4.9	Representative terms selected by LSI and Log-Likelihood Ratio Tests for the largest three clusters in Project Demo 1 on terrorism research (1996–2003)	164
Table 4.10	The popularity of a few terms on Google as of July 26, 2017.	169
Table 4.11	The information entropy of a term in articles of terrorism research (1996–2003).	171
Table 4.12	Top 20 content-bearing terms from the largest cluster of terrorism research (1990–2003) along with top terms identified by TF*IDF.	174
Table 4.13	Factors that may influence citations of a scientific publication.	187
Table 4.14	Information stored in the Articles table of the wos database.	192
Table 4.15	The same article is indexed differently in different sources.	193
Table 4.16	The number of articles distributed in subject categories.	194
Table 4.17	Top 20 keywords associated with papers from the Engineering subject category	195
Table 4.18	The average number of citations per paper and that of references per paper are both field-dependent	198
Table 5.1	Major fields and illustrative values	214
Table 5.2	Output files from SemRep	217
Table 6.1	A list of well-known, open source NLP tools	224
Table 6.2	The sample relation types in the news domain	230
Table 6.3	Concept table.	237
Table 6.4	CONCEPT_SEMTYPE table.	237
Table 6.5	PREDICATION table	237
Table 6.6	PREDICATION_ARGUMENT table.	238
Table 6.7	The list of related terms to “farm”.	242
Table 6.8	The list of core options available in MALLET	244
Table 6.9	The number of topics and top terms generated by LDA.	245
Table 6.10	Number of POS, words, Synsets, and sense pairs	249
Table 6.11	Polysemy information	249
Table 6.12	Statistics of BabelNet 3.7	251
Table 6.13	Options available for ID retrieval in BabelNet API.	252
Table 6.14	The results of the ID retrieval REST API	252

Table 6.15	Options available for word sense retrieval of BabelNet API	253
Table 7.1	Classification of the biologically meaningful verb list	278
Table 7.2	A list of strategies that characterize relation between two entities	278
Table 7.3	Example of output of extracted entities	279
Table 7.4	Example of output of extracted relations	279
Table 8.1	The complexity of a query can be reduced in a graph database	285
Table 8.2	Questions and corresponding queries in Cypher to a graph of scientific publications	288
Table 8.3	The graph constructed from the semantic predications on virus	291
Table 8.4	Causality claims concerning Ebola	291
Table 8.5	Papers that made the earliest causality claims on Ebola	292
Table 8.6	Negations of causality claims on Ebola	293
Table 8.7	MEDLINE articles that hypothesized causal relations in titles	294
Table 8.8	MEDLINE articles that hypothesized causal relations in abstracts.	295
Table 8.9	Earliest sentences concerning the predication: Virus CAUSES Neoplasm (PID: 544471)	296
Table 8.10	Semantic predications with burstness strengths > 10.0 from top 167 ones with bursts. ‘=>’ denotes CAUSES	298
Table 8.11	Top 10 most frequently appeared semantic predications in the virus subset	303
Table 8.12	Some examples of rare predications with 1, 5, or 10 appearances in total	303
Table 8.13	Statistics of semantic predications concerning viruses	305
Table 8.14	Top 20 most popular types of predicates in the virus dataset	305
Table 8.15	Most frequent terms mapped to the UMLS Concept HIV (CUI = C0019682) as subjects and objects.	310
Table 8.16	Semantic predications on causal relations from the 1983 article (PubMed ID: 6870184).	320
Table 8.17	All the semantic predications associated with the MEDLINE article (PMID: 14766405)	324
Table 8.18	Two major semantic predications in cluster #0	328

Table 8.19	Source sentences of the HHV8 and KS predication in articles published in 1996	329
Table 8.20	Semantic predications extracted from the article with five transformative links (PMID: 9989205)	332
Table 8.21	Semantic predications extracted from article PMID: 1134302.	335
Table 9.1	Sentences that indicate uncertainties in scientific knowledge	342
Table 9.2	The uncertainties of scientific disciplines	344
Table 9.3	A seed list of uncertainty cue words	346
Table 9.4	Semantic predications extracted from Nakao et al. (PMID: 12531578)	352
Table 9.5	Distributions of uncertainty cue words in scientific publications (with reference to the term knowledge)	354
Table 9.6	Distributions of uncertainty cue words in non-scientific publication sources	355
Table 9.7	12 dialectical relations identified by Murray Davis	356
Table 9.8	HIV causes AIDS (with the green background) and HIV is not the cause of AIDS (with the pink background).	357
Table 9.9	The knowledge of the cause of dementia in patients with AIDS	359
Table 9.10	Purposes served by meta-discourse	361
Table 9.11	Examples of claims and leading meta-discourse	363
Table 9.12	Sentences containing the word 'uncertainty' in MEDLINE articles.	363