



Massachusetts Institute of Technology
Engineering Systems Division

ESD Working Paper Series

Exploring Terms and Taxonomies Relating to the Cyber International Relations Research Field: or are "Cyberspace" and "Cyber Space" the same?

Steven Camiña

Massachusetts Institute of Technology
77 Massachusetts Avenue
Cambridge, MA 02139

Nazli Choucri

Massachusetts Institute of Technology
77 Massachusetts Avenue
Cambridge, MA 02139

Stuart Madnick

Massachusetts Institute of Technology
77 Massachusetts Avenue
Cambridge, MA 02139

Wei Lee Woon

Masdar Institute of Science and
Technology



**Exploring Terms and Taxonomies Relating to
the Cyber International Relations Research Field:
or are "Cyberspace" and "Cyber Space" the same?**

Steven Camiña
Stuart Madnick
Nazli Choucri
Wei Lee Woon

Working Paper CISL# 2011-03

August 2011

Composite Information Systems Laboratory (CISL)
Sloan School of Management, Room E62-422
Massachusetts Institute of Technology
Cambridge, MA 02142

Exploring Terms and Taxonomies Relating to the Cyber International Relations Research Field: or are "Cyberspace" and "Cyber Space" the same?

Steven Camiña, Stuart Madnick, Nazli Choucri
Massachusetts Institute of Technology

Wei Lee Woon
Masdar Institute of Science and Technology

August 2011

ABSTRACT

This project has at least two facets to it: (1) advancing the algorithms in the sub-field of bibliometrics often referred to as "text mining" whereby hundreds of thousands of documents (such as journal articles) are scanned and relationships amongst words and phrases are established and (2) applying these tools in support of the Explorations in Cyber International Relations (ECIR) research effort. In international relations, it is important that all the parties understand each other. Although dictionaries, glossaries, and other sources tell you what words/phrases are supposed to mean (somewhat complicated by the fact that they often contradict each other), they do not tell you how people are actually using them.

As an example, when we started, we assumed that "cyberspace" and "cyber space" were essentially the same word with just a minor variation in punctuation (i.e., the space, or lack thereof, between "cyber" and "space") and that the choice of the punctuation was a rather random occurrence. With that assumption in mind, we would expect that the taxonomies that would be constructed by our algorithms using "cyberspace" and "cyber space" as seed terms would be basically the same. As it turned out, they were quite different, both in overall shape and groupings within the taxonomy.

Since the overall field of cyber international relations is so new, understanding the field and how people think about (as evidenced by their actual usage of terminology, and how usage changes over time) is an important goal as part of the overall ECIR project.

1. INTRODUCTION

This paper is an extension of the work in [Camina 2010] that investigates the modeling of research landscapes through the automatic generation of hierarchical structures (taxonomies) comprised of terms related to a given research field. Taxonomy generation algorithms are based on the analysis of a data set of bibliometric information obtained from a credible academic online publication database. In particular, this paper analyzes the online publication databases within Engineering Village, namely Compendex and Inspec, by querying them using the query terms (*seed terms*) such as "cyber", "cyberspace", "cyber space", and "internet,"

1.1 Sources Used

Engineering Village¹ is a combination of three online databases: Compendex, Inspec and NTIS. Compendex and Inspec are both significantly larger in scope compared to NTIS (National Technical Information Service). The latter is a database of government reports and information covering several product categories ranging from administration/management to earth sciences. Because of NTIS's limited scope compared to Compendex and Inspec, we focused our data gathering efforts on Compendex and Inspec. Compendex and Inspec cover publications from 1884 up to the present and are available free of charge to members of the MIT community, allowing our research group to query the online publication database as often as we wanted without any overhead.

Compendex is a comprehensive bibliographic database of scientific and technical engineering research, covering all engineering disciplines. It includes millions of bibliographic citations and abstracts from thousands of engineering journals and conference proceedings. Compendex covers well over 120 years of core engineering literature. Specifically, Compendex includes over 5 million summaries of journal articles and conference proceedings and 220,000 new additions every year. Over 5,000 engineering journals and conferences are indexed and the database is updated weekly. Coverage of Compendex includes: Mechanical Engineering, Civil Engineering, Electrical Engineering and Electronics, Chemical Engineering and Aeronautical Engineering. Compendex is produced by Elsevier Engineering Information Inc.

Inspec includes bibliographic citations and indexed abstracts from publications in the fields of physics, electrical and electronic engineering, communications, computer science, control engineering, information technology, manufacturing and mechanical engineering, operations research, material science, oceanography, engineering mathematics, nuclear engineering, environmental science, geophysics, nanotechnology, biomedical technology and biophysics. Inspec contains over eight million bibliographic records taken from 3,000 scientific and technical journals and 2,000 conference proceedings. Over 400,000 new records are added to the database annually. Online coverage is from 1969 to the present, and records are updated weekly. Inspec is produced by the Institution of Engineering and Technology (IET).

1.2 Data Obtained by Querying the Sources

Querying each database using the seed terms produces results which are a set of documents related to the seed term. Figure 1 shows a screenshot of a results page in Engineering Village for the search term "renewable energy."

¹ Available via www.engineeringvillage.com

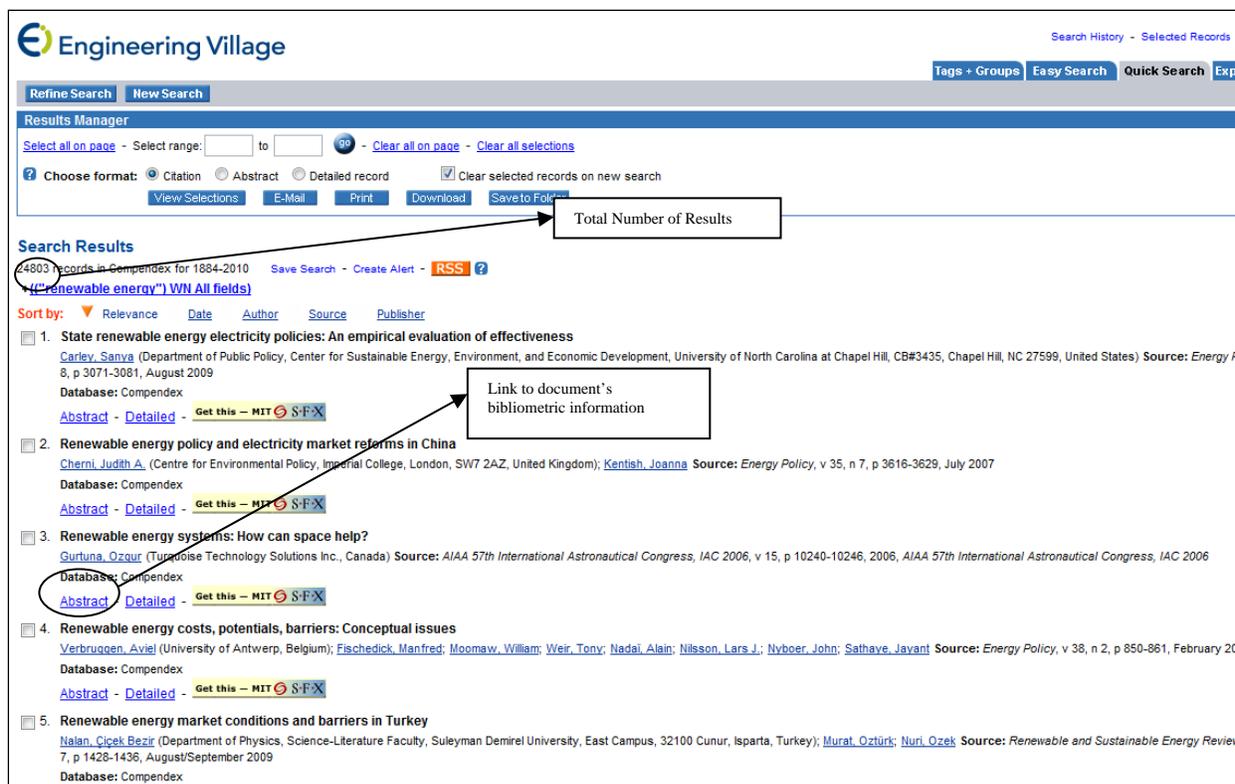


Figure 1: Screenshot of Results Page in Engineering Village. Highlighted within the figure are the locations in the website where the total number of results are shown and a link to a document’s bibliometric information

The software we developed then extracts (often referred to as “scraping”) each document’s bibliometric information from the website. Specifically, we took the document’s title, abstract, and keywords. Keywords within the document came in two varieties, controlled and uncontrolled. Each document has multiple controlled and uncontrolled keywords, which we refer to as the *terms* of each document. For a detailed description of the process used to gather bibliometric information and store keywords, please refer to Chapter 3 of [Camina 2010].

The bibliometric information of the various articles scraped from online publication databases is then stored into a local file (in SQLite3 format), which can then be manipulated as without needing to access the online publication database again. We refer to the collection of documents stored in the local file as the *data set* of bibliometric information. With the data set on hand, the rest of the analysis can be done without the need of an internet connection. Using the data set it is possible to:

1. Analyze all the keywords, which we refer to as *terms*, within all the documents in the data set.
2. Take the terms and generate a taxonomy, which is a hierarchical organization of the terms.

Unfortunately, the online interface of Engineering Village has a slight downside in that it only allows the user to view 4,025 documents at a time. In Figure 1, there are 24,803 results / documents for the seed term “renewable energy”, however, the online interface of Engineering Village only permits the browsing of the first 4,025 documents. There is a workaround for this, however, that is time-intensive and

involves non-automated steps. As such, for the analysis described in this paper, only the first 4,025 most relevant documents that came up in the search query results are considered.

In gathering the results using the seed terms mentioned previously, either Compendex or Inspec was used by querying each database using the seed term and seeing which database generated more results. The one that had more results is the one chosen to gather bibliometric information from.

2. RESULTS

2.1 Choosing which of Compendex / Inspec to use to gather bibliometric information from

Seed Term	Compendex Document Count	Inspec Document Count
“cyber”	5,293	4,096
“cyberspace”	983	637
“cyber space”	968	720
“internet”	117,394	21,317

Table 1: Result Counts for Seed Term Queries to Compendex and Inspec

Based on the results shown in Table 1 above, it can be seen that Compendex is the better online publication database to use when collecting bibliometric information related to “cyber”, “cyberspace”, “cyber space”, and “internet”. It must be noted that for “cyber” and “internet”, only the first 4,025 most relevant documents were taken into consideration. It must also be noted that despite attempting to store all the 4,025 documents, there inevitably are several instances where the document’s data cannot be gathered for some reason – either an unexpected error in the website or some abnormal textual (ASCII) representation of data. As such, the final data set size is slightly less than the original document counts displayed in the Compendex online interface.

2.2 Terms in Each Data Set

Table 2 summarizes the number of terms contained in each data set generated by a particular seed term.

Seed Term Used to Generate Data Set	Total Number of Terms in Data Set
“cyber”	14,893
“cyberspace”	3,488
“cyber space”	4,717
“internet”	14,734

Table 2: Summary of Terms Contained in Each Data Set

2.3 Common Terms Between Data Sets

In Table 3 below, the data set generated using the seed term in the first column is compared to the data set generated using the seed term in the second column and the number of common terms is found.

Seed Term Used to Generate Data Set	Seed Term Used to Generate Data Set	Number of Terms in Common
“cyber”	“cyberspace”	2,049
“cyber”	“cyber space”	4,218
“cyber”	“internet”	3,812
“cyberspace”	“cyber space”	1,338
“cyberspace”	“internet”	1,511
“cyber space”	“internet”	1,659

Table 3: Number of Terms in Common Between Data Sets

2.4 Determining Percentage Similarity Between Data Sets

Determining an accurate value for percentage similarity of terms between data sets was tricky because each data set had a different number of terms contained within it. For example, if a pair of data sets had 1,000 terms in common but one data set had 1,500 terms total and the other had 1,000,000 terms total, then from one data set’s perspective, the overlap was significant, but from the other it seems trivial. In order to avoid this confusion, we decided to take rank the terms within each data set according to frequency of occurrence within documents, and then compare the top X terms in one data set to the top X terms in another.

Frequency of occurrence of terms within documents is determined by counting how many times the *stem* of a term occurs as one of the keywords within a document. For example, if a document collected from an online publication database has the keywords: [“information”, “browser”, “security”], while another document has [“service provider”, “browsers”, “government control”], the term / keyword “browser” will be counted as occurring in both documents, as the terms “browser” and “browsers” have the same stem. For a more detailed description of word stemming and keyword / term collection, please refer to [Camina 2010]. Figures 2 and 3 summarize the results.

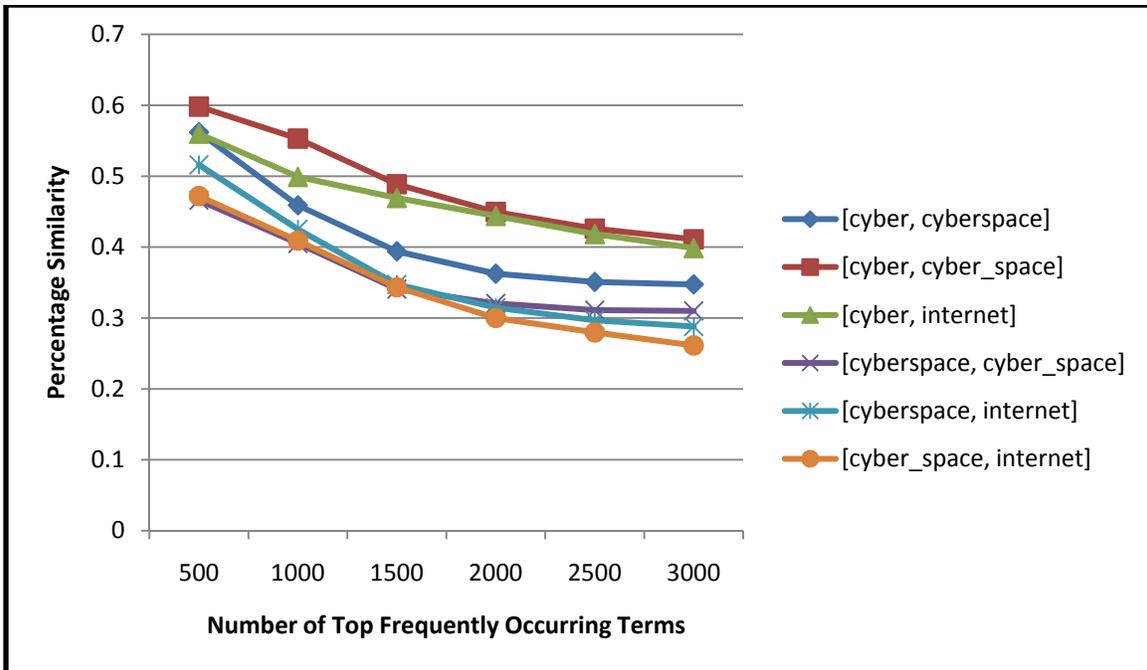


Figure 2: Percentage Similarity between terms in each of the data sets using the top 500-3000 most frequently occurring terms in the data sets

The figure below is similar to figure 2 above except for the term values range and granularity.

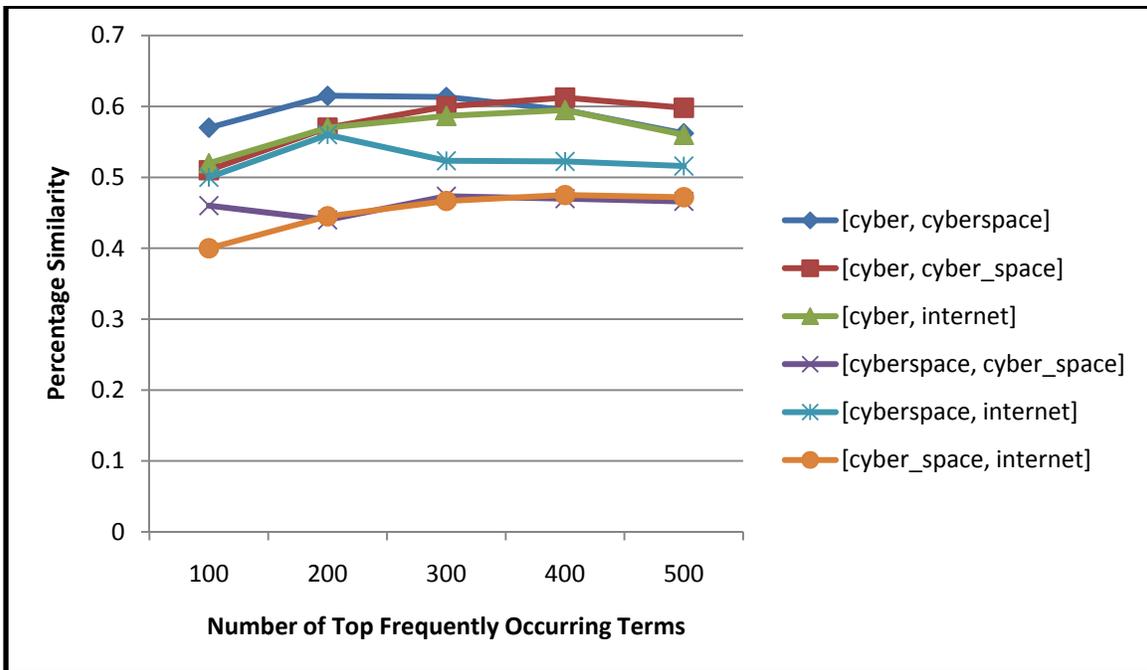


Figure 3: Percentage Similarity between terms in each of the data sets using the top 100-500 most frequently occurring terms in the data sets

One key observation was that for large values of top frequently occurring terms used (500-3000), there is a general decreasing trend of percentage similarity as the value for the number of top frequently occurring terms is increased. However, for smaller values of top frequently occurring terms (100-500), the opposite is true. This shows that the majority of the common terms between data sets happen among the most frequently occurring terms.

2.5 Terms In Common Across All Databases

The following 886 terms were found to be common across all four data sets that we generated. These are listed alphabetically below. The terms represent the concepts within the “cyberspace” research landscape.

abstract modeling	broadcasting	computer aided software	context-aware services
abstracting	calculations	engineering	contracts
access control	calibration	computer animations	control
accidents	cameras	computer architecture	control of networks
acoustic signal processing	campus network	computer crime	control systems
acoustics	cellular phones	computer forensics	control theory
ad hoc networks	cellular telephone systems	computer games	convergence (mathematics)
adaptive algorithms	channel capacity	computer graphics	convergence of numerical
adaptive control systems	chaos theory	computer hardware	methods
adaptive systems	character recognition	computer hardware	copying
administrative data	charge coupled devices	description languages	copyrights
processing	china	computer integrated	correlation methods
agents	classification (of	manufacturing	cost benefit analysis
agglomeration	information)	computer monitors	cost effectiveness
alarm systems	client server computer	computer music	cost reduction
algorithms	systems	computer networks	costs
america	closed loop control systems	computer operating systems	critical component
amplitude modulated	cluster head	computer privacy	cross-cultural study
analytical models	clustering	computer programming	cryptographic algorithms
animated movies	codes (standards)	languages	cryptography
animation	codes, symbolic	computer programs	cultural difference
antennas	coding errors	computer securities	current technology
anthropometry	cognitive systems	computer simulation	current trends
applications	collaboration systems	computer simulation	curricula
apriori	collaborative designs	languages	customer loyalty
arsenic compounds	collaborative filtering	computer software	customer satisfaction
artificial intelligent	collaborative work	computer software	cyber communities
artificial life	collision avoidance	reusability	cyber crimes
audio acoustics	color	computer supported	cyber spaces
audio systems	color image processing	cooperative work	cybercrime
audition	combinatorial mathematics	computer system firewalls	cybernetics
augmented reality	command and control	computer systems	cyberspace
authentication	systems	computer systems	data acquisition
automata theory	commerce	programming	data communication
automatic generation	communication	computer technology	systems
automation	communication channels	computer viruses	data compression
autonomous agents	(information theory)	computer vision	data handling
autonomous behaviors	communication overheads	computer worms	data integrations
bandwidth	communication sessions	computers	data managements
basic theory	communications systems	computers - applications	data privacy
bayesian network	competition	conceptual frameworks	data processing
behavior modeling	complex systems	concurrency control	data recording
behavioral research	computational complexity	concurrent engineering	data reduction
benchmarking	computational geometry	conformal mapping	data securities
best efforts	computational intelligence	congestion control	data sets
biology	computational linguistics	(communication)	data storage
bipartite graphs	computational methods	constraint theory	data storage equipment
blind source separation	computational science	consumer electronics	data structures
blogospheres	computer aided design	content based retrieval	data transfer
broadband networks	computer aided instruction	context information	data visualization
broadband services		context-aware	database systems

data-mining
 ddos attacks
 decision making
 decision support systems
 decision supports
 decision theory
 decoding
 decomposition
 degrees of freedom
 (mechanics)
 denial of service attack
 department of defense
 design
 design elements
 design method
 detection system
 developing countries
 different mechanisms
 digital arithmetic
 digital communication
 systems
 digital content
 digital convergence
 digital files
 digital formats
 digital image
 digital image storage
 digital informations
 digital libraries
 digital media
 digital signal processing
 digital technologies
 digital television
 digital watermarking
 disaster prevention
 disks (structural
 components)
 display devices
 distance education
 distributed computer
 systems
 distributed parameter
 control systems
 distributed processing
 domain knowledge
 dynamics
 dynamics analysis
 e sciences
 e-business
 echo suppression
 ecology
 e-commerce
 economic analysis
 economics
 education
 education computing
 educational institutions
 efficient method
 eirev
 e-learning
 electric breakdown
 electric network analysis
 electric network topology
 electromagnetic waves
 electromagnetism
 electronic commerce
 electronic communication
 electronic data
 electronic data interchange
 electronic document
 identification systems
 electronic mail
 electronic publishing
 electronic transaction
 electronic warfare
 electronics industry
 electronics packaging
 elsevier (co)
 embedded systems
 emergency responses
 emerging technologies
 empirical research
 employment
 enabling technologies
 encoding (symbols)
 end-users
 energy-efficient
 engineering education
 engineering research
 english languages
 enterprise computing
 entropy
 environmental conditions
 environmental engineering
 environmental impact
 environmental monitoring
 environmental protection
 error analysis
 error correction
 error detection
 e-services
 e-trading
 evaluation
 evolutionary algorithms
 existing methods
 experiments
 expert systems
 face recognition
 facial animation
 facsimile
 factory automation
 fast fourier transforms
 feature extraction
 feature selections
 feedback controller
 feedbacks]
 fiber optic networks
 file sizes
 file system
 finance
 finite automata
 flow interactions
 force feedback
 forecasting
 formal languages
 formal logic
 formal models
 fractals
 function evaluation
 functions
 fuzzy logic
 fuzzy sets
 game theory
 gateways (computer
 networks)
 general (co)
 genetic algorithms
 geographical information
 system
 geostationary satellites
 gesture recognition
 global networks
 global optimizing
 global positioning system
 graph theory
 graphic methods
 graphical user interfaces
 grid computing
 group communication (gc)
 groupware
 handicapped persons
 haptic interfaces
 hard disk storage
 harvesting
 hazards
 health
 health care
 health information
 hearing-impaired
 hearing-impaired users
 heidelberg (co)
 hierarchical systems
 high definition television
 higher educations
 holography
 hospital information
 systems
 hospitals
 html
 human behaviors
 human computer
 interaction
 human engineering
 human factors
 hypertexts
 identification (control
 systems)
 identity theft
 image analysis
 image coding
 image communication
 systems
 image compression
 image data
 image enhancement
 image processing
 image quality
 image reconstruction
 image retrieval
 image segmentation
 imaging systems
 imaging techniques
 independence (personality)
 independent variables
 in-depth interviews
 indexing
 individual (pss 544-7)
 industrial applications
 industrial economics
 industrial engineering
 industrial management
 information analysis
 information and
 communication
 technologies
 information dissemination
 information exchanges
 information fusion
 information infrastructures
 information management
 information networking
 information privacy
 information processing
 information retrieval
 information retrieval
 systems
 information science
 information security
 information services
 information systems
 information technology
 information theory
 information use
 information visualization
 innovation
 innovation processes
 innovative solutions
 integrated circuit layout
 integrated sources
 integrity
 intellectual property
 intelligent agents
 intelligent buildings
 intelligent control
 intelligent networks
 intelligent robotics
 intelligent services
 intelligent systems
 intelligent vehicle highway
 systems
 interactive computer
 graphics
 interactive computer
 systems
 interactivity
 interconnection networks
 interface designs
 interfaces (computer)
 international conferences
 international cooperation
 international law
 international trade
 internet
 internet cafe
 internet gamings
 internet monitoring
 internet protocol
 internet protocol (ip)
 internet protocol networks
 internet services
 internet technology
 internet use
 internet users
 internet2
 internet-2
 interoperable
 intrusion detection
 intrusion-detection systems
 investments
 issues and challenges
 it security
 java programming language
 kalman filtering
 ketones
 key performance indicators
 key problems
 know-how
 knowledge acquisition
 knowledge based systems

knowledge engineering	monitoring system	outsourcing	regulatory compliance
knowledge management	motion picture experts	p2p system	reliability analysis
knowledge representation	group (mpeg)	packet networks	reliable
landforms	motion picture experts	paradigm shifts	remote control
language processing	group standards	parallel processing systems	remote education
large scale systems	motion pictures	parameterization	remote sensing
laws and legislation	multi dimensional	patient monitoring	remote users
learning algorithms	multi-agent	pattern matching	research
learning systems	multi-agent system	peer-to-peer networks	research activities
legal frameworks	multicasting	performance	research and development
level of details	multi-hop communications	personal computers	management
libraries	multimedia services	personal digital assistants	research communities
life-cycle	multimedia systems	personal information	research results
light	multiple cameras	personalization	residual energy
light measurement	multiple sources	personnel	resource allocation
linguistics	multiplexing	personnel training	resource sharing
linux - operating systems	museums	pervasive computing	response time (computer
liquid crystal displays	nanotechnologies	petri nets	systems)
local area networks	nash equilibrium	philosophical aspects	retransmissions
logic programming	natural frequencies	photography	reusability
low costs	natural resources	photons	revenue
low-power	navigation	physical world	rfid technology
machine design	negative impacts	pixels	risk analysis
machine-learning	network architecture	plain text	risk assessment
malicious activities	network attacks	planning	risk management
malicious software	network intrusion	policy-makers	road maps
man machine systems	detections	polynomials	roads and streets
management	network management	portable equipment	robotics
management - information	network monitoring	portals	role-playing games
systems	network operators	printing	rom
management systems	network protocols	printing presses	rough set theory
manipulators	network resources	probability	routers
maps	network size	probability density function	routing protocols
marketing	network technologies	probability distributions	sales
markov processes	network topology	problem oriented languages	satellite communication
mathematical models	network traffic	problem solving	systems
mathematical	networked systems	process control	scada systems
transformations	networks security	process information	scheduling
matrix algebra	neural networks	product design	school buildings
maximum likelihood	new approaches	production control	search engines
estimation	new concept	productivity	security infrastructures
medical applications	next generation networks	profile	security levels
medical computing	normal-hearing (nh)	program processors	security management
medical imaging	novel methods	programmable logic	security mechanisms
medical records	numerical methods	controllers	security of data
message passing	object oriented	project management	security protocols
metadata	programming	proof of concepts	security requirements
microwave antennas	object recognition	protocol designs	security services
microwaves	ocean engineering	prototype implementation	security situation
military applications	oceanography	prototype system	security systems
military communications	offline	public key cryptography	self-organize
military operations	online communities	public policy	semantic information
mining	online conferencing	public space	semantic web
mobile ad hoc networks	online discussions	qos requirements	semantics
mobile ad hoc networks	online forum	quality assurance	semiconductor quantum
(manet)	on-line gamings	quality control	dots
mobile agents	online learning	quality of service	sensor data fusion
mobile computers	online shopping	quantitative method	sensor fusion
mobile devices	online systems	query languages	sensor networks
mobile nodes	ontology	radio broadcasting	sensor nodes
mobile phones	open source software	radio communication	sensors
mobile robots	open sources	random processes	sensory perception
mobile telecommunication	open systems	real time systems	servers
systems	operational modeling	real times	service discovery
mobile users	operational systems	real-space	service provider
modal analysis	optical communication	real-world	service quality
models	optical data processing	reasoning process	service users
modernization	optical fibers	recommendation systems	set theory
modulation	optical systems	redundancy	sign language
monitoring	optimization	regression analysis	signal detection

signal encoding	springs (components)	television broadcasting	video signal processing
signal filtering and prediction	spurious signal noise	temporal pattern	video telephone equipment
signal interference	standardization	testing	videodisks
signal processing	state of the art	text processing	videotex
signal receivers	statistical features	theorem proving	virtual communications
signal theory	statistical methods	theoretical models	virtual environment
signal to noise ratio	storage spaces	three dimensional	virtual reality
simulation experiments	strategic planning	three dimensional computer graphics	virtual spaces
simulation results	structural analysis	time and space	virtual worlds
simulator models	structural characteristics	time complexity	virtualizations
simulators	students	tools	visual communication
single machines	supercomputer	tools and techniques	visualization
situation awareness (sa)	surveillance	topology	voice/data communication systems
situation-awareness	surveys	trace analysis	wavelet transforms
smart cards	synchronization	traceback	web 2.0
smart devices	syntactics	tracking (position)	web applications
social contexts	system architecture	transaction cost	web browsers
social environment	system monitoring	transcoding	web impact factor(wif)
social issues	system use	translation (languages)	web information
social networking	systems analysis	transmission control protocol	web intelligence
societies and institutions	systems engineering	two-dimension	web interfaces
socioeconomic status (ses)	teaching	ubiquitous computing	web pages
software agents	technical development	ubiquitous networks	web portals
software architecture	technical presentations	uncertainty analysis	web servers
software engineering	technological forecasting	undergraduate students	web service
software prototyping	technological solutions	underwater acoustics	websites
software-based	technology	upper bound	wide area networks
source codes	technology transfer	use cases	wireless communications
south korea	technology-based	user activities	wireless networks
space platforms	tele immersion	user experience	wireless sensor networks
space research	telecommunication	user interfaces	wireless telecommunication systems
special effects	telecommunication equipment	user networks	word processing
specifications	telecommunication links	user preferences	work environments
spectrum analysis	telecommunication networks	user requirements	work in progress
speech	telecommunication services	user-centric	workplace
speech analysis	telecommunication systems	variational techniques	world wide web
speech coding	telecommunication traffic	vector quantization	xml
speech communication	teleconferencing	vectorization	(e ,2e) theory
speech intelligibility	telegraph	vehicles	(e ,3e) process
speech processing	telephone	video cameras	(i ,j) conditions
speech recognition	telephone systems	video conferencing	(otdr) technology
speech synthesis	telepresence	video contents	(r ,s ,s) policy
speech transmission	television	video recording	
springer (co)			

2.6 Taxonomy Generation

The next step was compare the taxonomies generated using the 886 terms in common mentioned above. To do this, we used the 886 terms above as the term list of the taxonomy, and used each of the data sets gathered as backend for the taxonomy generation algorithms described in Chapter 3 of [Camina 2010]. Two sets of taxonomies were generated for each data set, each using a different algorithm. The two algorithms represent the best taxonomy generation algorithms as motivated and described in [Camina 2010]. These algorithms are:

1. Heymann algorithm, closeness centrality, cosine similarity metric (H-CC)
2. DJP algorithm, asymmetric NGD similarity metric, closeness centrality for root selection (D-SC)

2.6.1 Root Terms In Taxonomies Generated

In our implementation of taxonomy generation, the *seed term* used to generate the data set is not the same as the *root term*, or term at the top of the hierarchy in the taxonomy generated. The choice as to which term becomes the generated root term is dependent upon the centrality of the term in the distance matrix, which is an abstract representation of the data set. For a more detailed description of the distance matrix and the term similarity metrics used to construct it, please refer to [Camina 2010].

Table 4 summarizes the root terms found for each taxonomy generated using the two algorithms mentioned previously.

Seed Term Used to Generate Data Set	HCC Root Term	DSC Root Term
“cyber”	Computers.	Cyber Spaces.
“cyberspace”	Cyberspace.	Wireless Sensor Networks.
“cyber space”	Computers.	E-Sciences.
“internet”	Internet.	Visualization.

Table 4: Root Terms For Each Taxonomy Generated

Note that taxonomies generated using a different taxonomy generation algorithm or a different backend data set are different not just in the root term of the taxonomy but in many of the term links as well.

It must also be noted that the correctness of root terms is improves as the size of the backend data set increases. Based on analysis in [Camina 2010], the ideal data set size is in the 10^5 magnitude range, however the size of the data sets used to generate the taxonomies in our analysis is only in the 10^3 to 10^4 range.

2.6.2 Comparison of Taxonomies Generated

2.6.2.1 Using the H-CC algorithm for Taxonomy Generation

Table 5 below shows pairwise comparisons between each of the four taxonomies generated using the H-CC algorithm. The first two columns indicate the taxonomies compared and the third column shows the percentage similarity within the links of the taxonomies. Note that since the two taxonomies compared both use the same term list (the 886 term list shown previously), the taxonomies are directly comparable. Taxonomies are compared by calculating the number of similar links they share as a percentage of the total number of links in the taxonomy.

Seed Term Used to Generate Data Set that Serves as the Backend of the Taxonomy	Seed Term Used to Generate Data Set that Serves as the Backend of the Taxonomy	Percentage of Similar Links in Taxonomies Generated
“cyber”	“cyberspace”	19.64%
“cyber”	“cyber space”	30.47%
“cyber”	“internet”	24.72%
“cyberspace”	“cyber space”	19.41%
“cyberspace”	“internet”	15.24%
“cyber space”	“internet”	15.69%

Table 5: Percentage Similarity of Taxonomies Generated using H-CC algorithm

2.6.2.2 Using the D-SC algorithm for Taxonomy Generation

Table 6 below shows pairwise comparisons between each of the four taxonomies generated using the D-SC algorithm. The first two columns indicate the taxonomies compared and the third column shows the percentage similarity within the links of the taxonomies. Note that since the two taxonomies compared both use the same term list (the 886 term list shown previously), the taxonomies are directly comparable.

Seed Term Used to Generate Data Set that Serves as the Backend of the Taxonomy	Seed Term Used to Generate Data Set that Serves as the Backend of the Taxonomy	Percentage of Similar Links in Taxonomies Generated
“cyber”	“cyberspace”	11.40%
“cyber”	“cyber space”	19.07%
“cyber”	“internet”	10.05%
“cyberspace”	“cyber space”	9.82%
“cyberspace”	“internet”	6.66%
“cyber space”	“internet”	5.19%

Table 6: Percentage Similarity of Taxonomies Generated using D-SC algorithm

2.6.2.3 Comparing H-CC and D-SC Taxonomies

Table 7 compares the H-CC and D-SC taxonomies generated using the same backend data set.

Seed Term Used to Generate Data Set that Serves as the Backend of the Taxonomy	Percentage of Similar Links in H-CC and D-SC generated Taxonomies
“cyber”	29.91%
“cyberspace”	27.31%
“cyber space”	31.26%
“internet”	32.28%

Table 7: Comparison of H-CC and D-SC Taxonomies with Similar Backend Data Sets

2.6.3 Analysis of Taxonomies

Based on the information contained in Table 4 showing the root terms for each taxonomy generated, the most interesting looking taxonomies are the ones with the root terms: “computers”, “cyberspace”, and “internet”, corresponding to the following taxonomies:

1. “cyber” taxonomy generated using the H-CC algorithm
2. “cyberspace” taxonomy generated using the H-CC algorithm
3. “cyber space” taxonomy generated using the H-CC algorithm
4. “internet” taxonomy generated using the H-CC algorithm
5. “cyber” taxonomy generated using the D-SC algorithm

Each of these taxonomies are analyzed in the succeeding sections. Note that each of the taxonomies generated may use different backend data sets but they are all composed of the same terms. As such, each of the four taxonomies analyzed in the following sections have the same content but are just organized in four different ways.

High resolution copies of the GIF files for Figures 4-9 can be found and downloaded from <http://web.mit.edu/smadnick/www/ECIR/TaxonomyImages/> It is recommended that a flexible viewer be used, such as zgrviewer (from <http://zvtm.sourceforge.net/zgrviewer.html>).

2.6.3.1 “Cyber” Taxonomy Using H-CC Algorithm

Figure 4 shows a birds-eye view of the “cyber” taxonomy generated using the H-CC algorithm. For a closer view of the taxonomy, a GIF file of the taxonomy is available and is easily viewable with any default image viewer. The image can then be zoomed into for more granular inspection.

Some of the interesting observations about the taxonomy are listed below:

1. The root term of the taxonomy is “computers”
2. There are several interesting term clusters:
 - a. At the top of the taxonomy’s visualization, there is a cluster with “internet” as the root, leading to terms such as “internet use”, “internet protocol”, “email”, and “internet technology”
 - b. Underneath the “internet” cluster, there is another cluster with “computer crime” as the root, leading to terms such as “security systems”, “cyber crimes”, intrusion detection”, “computer forensics”, and “denial of service attacks”
 - c. Underneath the “computer crime” cluster, there is another cluster with “algorithms” as the root, leading to terms such as “optimization”, “learning algorithms”, and “adaptive algorithms”
 - d. Near the left-center of the taxonomy’s visualization, there is a cluster with “communication” as the root, leading to “telecommunication”, which in turn leads to terms such as “telecommunication networks”, “telecommunication services”, and “telephone”

- e. In the taxonomy, there is also a cluster with “speech” as the root, leading to terms such as “linguistics”, “speech recognition”, and “speech coding”

2.6.3.2 “Cyberspace” Taxonomy Using the H-CC Algorithm

Figure 5 shows a birds-eye view of the “cyber” taxonomy generated using the H-CC algorithm. For a closer view of the taxonomy, a GIF file of the taxonomy is available and is easily viewable with any default image viewer. The image can then be zoomed into for more granular inspection.

Some of the interesting observations about the taxonomy are listed below:

1. The root of the taxonomy is “cyberspace”
2. There is a cluster with “computers” as the root, leading to terms such as “computer crime”, “computer software”, “computer networks”, and “network security”
3. Similar to the taxonomy in “cyber” H-CC taxonomy in 2.3.6.1, there is a cluster with “internet” as the root

2.6.3.2 “Cyber space” Taxonomy Using H-CC Algorithm

Figure 6 shows a birds-eye view of the “cyber space” taxonomy generated using the H-CC algorithm. For a closer view of the taxonomy, a GIF file of the taxonomy is available and is easily viewable with any default image viewer. The image can then be zoomed into for more granular inspection.

Some of the interesting observations about the taxonomy are listed below:

1. The root term of the taxonomy is “computers”
2. Similar to the “cyber” cluster discussed previously in 2.6.3.1, this taxonomy also included the “telecommunication”, “speech” and “algorithms” clusters
3. In the taxonomy, there is a cluster with “technology” as the root, leading to terms such as “information technology”, “cyberspaces”, and “innovation”
4. In the taxonomy, there is a cluster with “disaster prevention” as the root, leading to terms such as “environmental impact”, and “security infrastructure”
5. There is a lot of noise / nonsense links in this taxonomy. In particular, there is a large cluster with “image enhancement” as the root, leading to several unrelated terms such as “identification”, “tracking”, “congestion control”, “internet protocol”, etc.

2.3.6.4 “Internet” Taxonomy Using the H-CC Algorithm

Figure 7 shows a birds-eye view of the “internet” taxonomy generated using the H-CC algorithm. For a closer view of the taxonomy, a GIF file of the taxonomy is available and is easily viewable with any default image viewer. The image can then be zoomed into for more granular inspection.

Some of the interesting observations about the taxonomy are listed below:

1. The root of the taxonomy is “internet”

2. Among the 4 taxonomies discussed in detail here, this taxonomy has the most shallow structure. It has a lot of terms at each level.
3. There are several interesting term clusters:
 - a. There is a cluster with “technology” as the root, leading to terms such as “internet technology”, “computer technology”, and “technology forecasting”. A similar cluster appeared in the “cyber space” taxonomy in 2.3.6.3, but the cluster described here is much larger
 - b. There is a cluster with “research” as the root, leading to terms such as “research and development management”, “behavioral research”, and “surveys”
 - c. There is a cluster with “semantics” as the root, located under “automation”, leading to terms such as “information theory”, “ontology”, “semantic web”, and “context-aware”
 - d. There is a cluster with “computers” as the root, leading to terms such as “computer crime”, “computer software”, “computer networks”, and “servers”
 - e. There is a cluster with “robotics” as the root, leading to terms such as “remote control”, “mobile robots”, and “intelligent robots”

2.3.6.5 “Cyber” Taxonomy Using D-SC Algorithm

Figure 8 shows a birds-eye view of the “cyberspace” taxonomy generated using the D-SC algorithm. For a closer view of the taxonomy, a GIF file of the taxonomy is available and is neasily viewable with any default image viewer. The image can then be zoomed into for more granular inspection.

Some of the interesting observations about the taxonomy are listed below:

1. The root of the taxonomy is “cyber spaces”
2. It is a very deep taxonomy, with only 2 terms in the first layer of terms in the taxonomy having child terms
3. There were no clear term clusters, however there were a few conceptual paths that could be traced. For instance, there was a path that had “cellular phones” → “cellular telephone systems” → “telephone systems” → “mobile phones”
4. In general, this taxonomy was much harder to read compared to the other three taxonomies discussed in this section

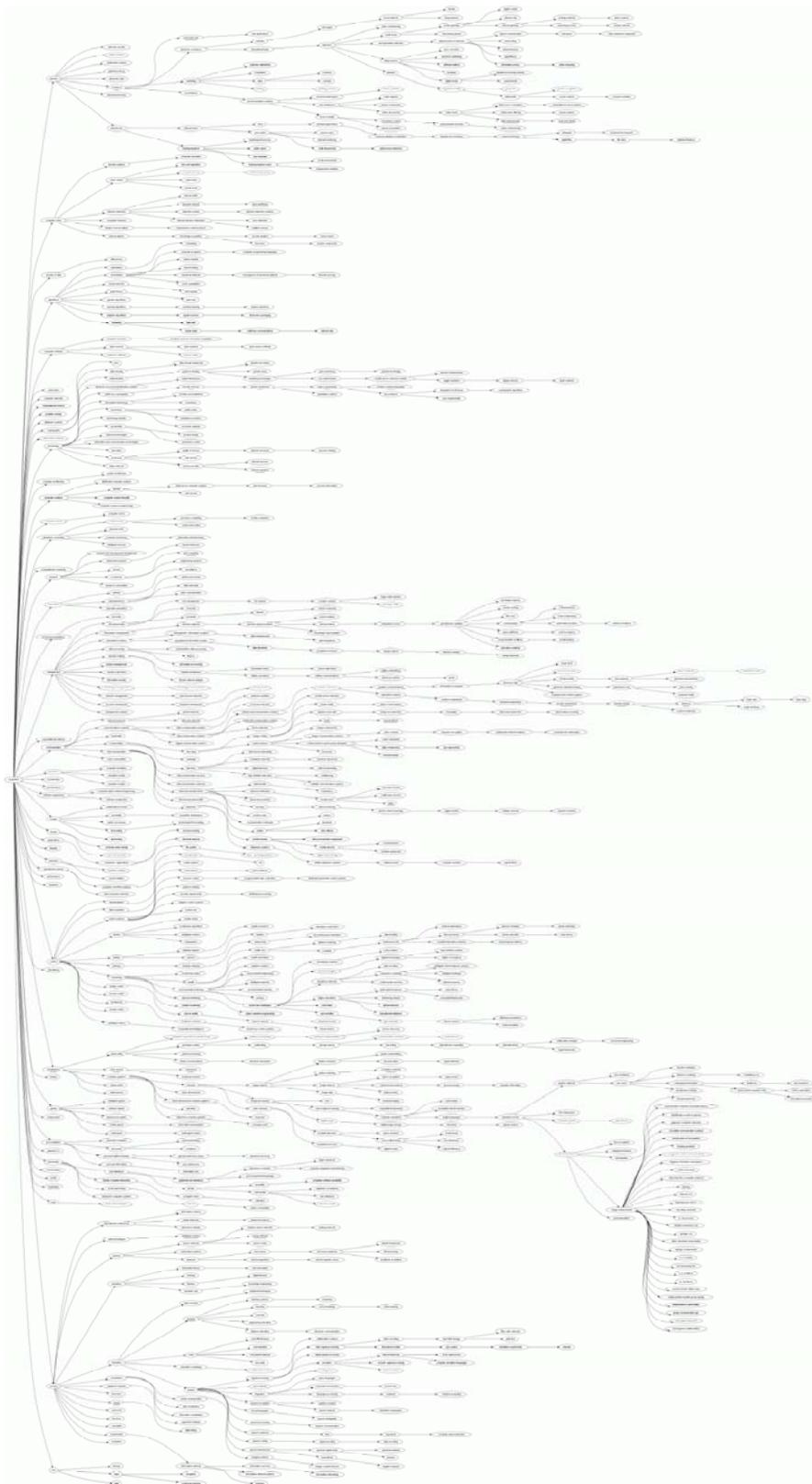


Figure 4: “Cyber” H-CC Taxonomy

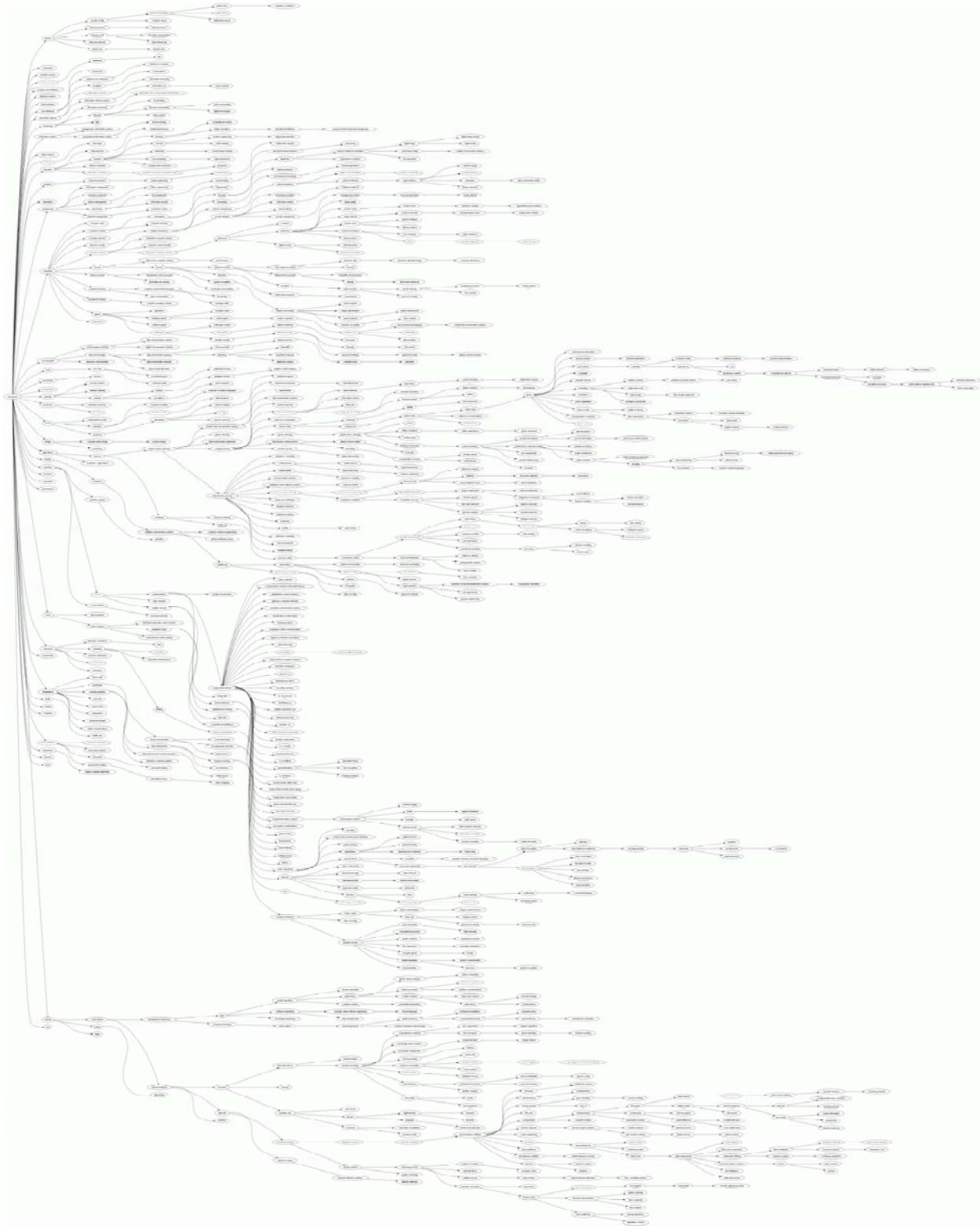


Figure 5: “Cyberspace” H-CC Taxonomy

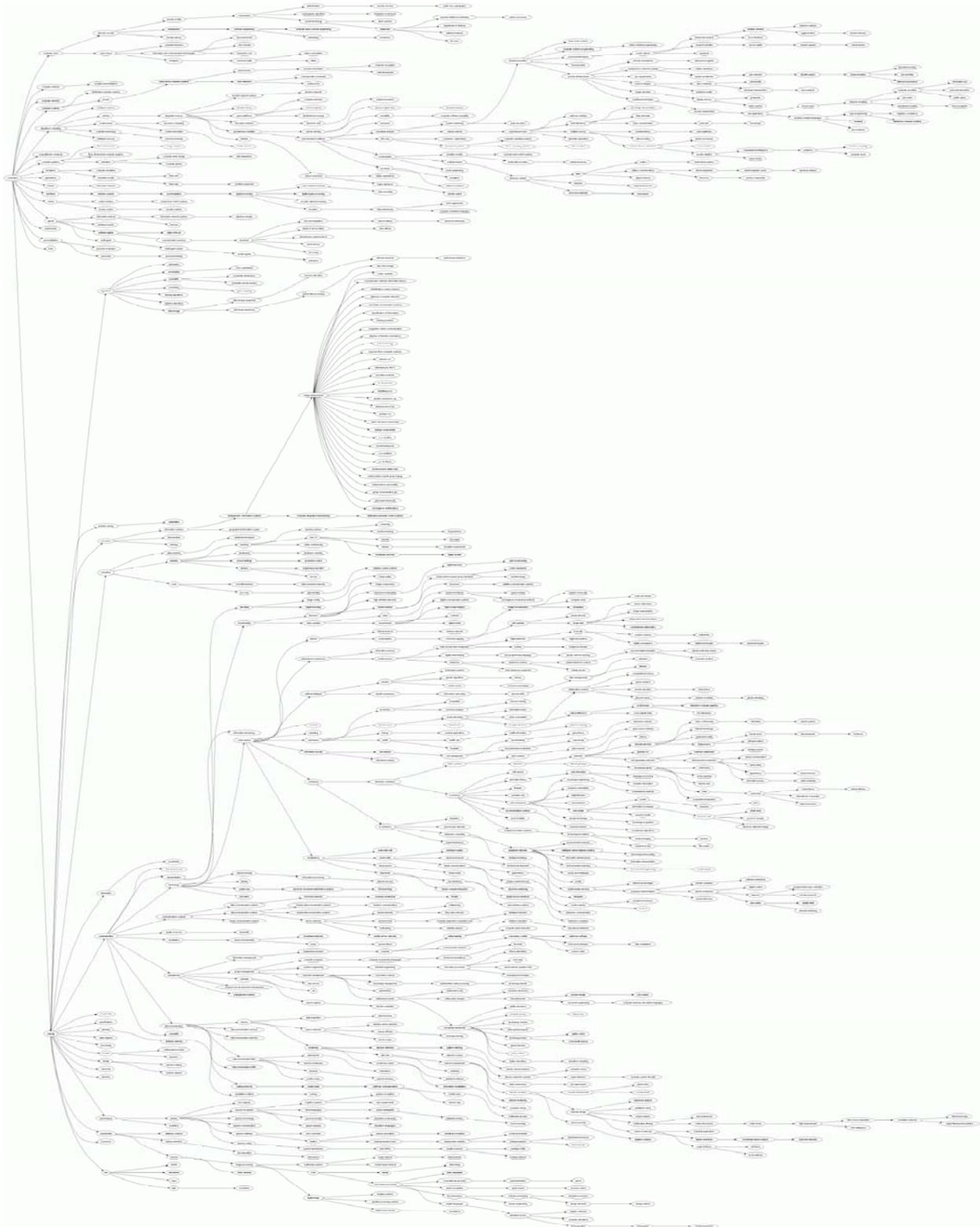


Figure 6: “Cyber space” H-CC Taxonomy

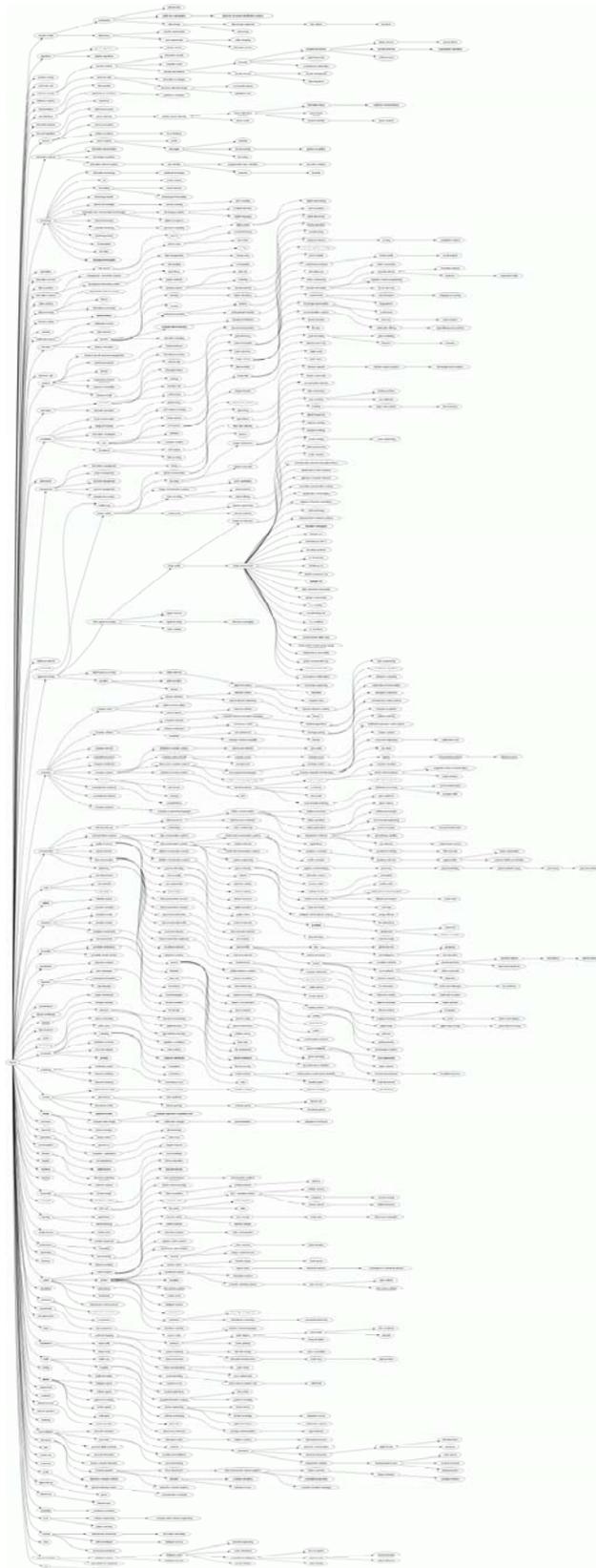


Figure 7: "Internet" H-CC Taxonomy

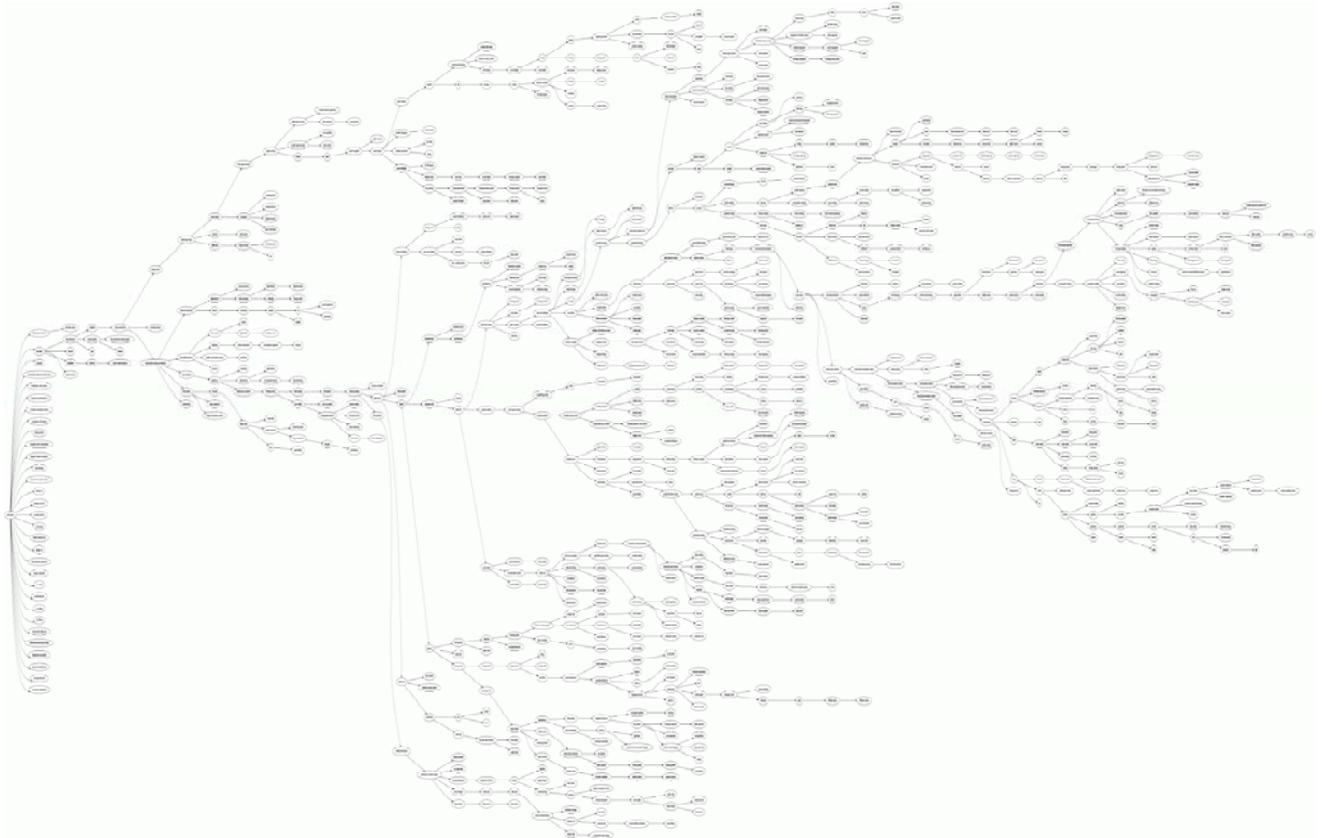


Figure 8: “Cyber” D-SC Taxonomy

3. CONCLUSIONS AND FUTURE RESEARCH

3.1 Are ‘cybersecurity’ and ‘cyber security’ the same?

Referring back to the sub-title, “are ‘cybersecurity’ and ‘cyber security’ the same? The results reported above indicate that there is definitely something different based upon the different taxonomies generated and displayed in Figures 5 and 6. The reasons for the differences are not immediately obvious – might be the ways that authors in different fields use the words (e.g., policy people vs. technology people), quirks of the algorithms, etc. That will be part of the future research that we intend to conduct, as well as other interesting directions listed below.

3.2 Future Research

This research raises almost as many issues as it answers, as noted in section 3.1 immediately above. Some areas of future investigation include:

3.2.1. Choice of type of sources: In this reported research, we have used academic publications. We could use blogs and news. What would that look like?

3.2.2. Choice of specific sources: How different are the taxonomies that are generated using different sources, such as Google Scholar, Scirus, Scopus, Web of Science, Engineering Village, etc as the pool of publications?

3.2.3. Choice of language: In this reported research, we have mainly focused on English publications, what if we included publications from other languages - probably translating the key words into English.

3.2.4. Finer grain source differences: What if we filtered the documents to separate them by region (what country they came from) or role (technology author vs policy author.) Would the taxonomies be similar or very different?

3.2.5. Temporal differences: How does the meaning and usage of terms, as represented by the taxonomy, change over time?

3.2.6. Algorithms: We have experimented with various algorithms for the automated generation of taxonomies. Which algorithms are best for our purposes?

3.2.7. Metric: What are the best ways to measure the quality of the algorithms and the results produced?

3.2.8. "Face validity": Would be good to show our automatically generated taxonomies to Subject Matter Experts (SMEs) to see whether they view the taxonomies as being meaningful.

ACKNOWLEDGEMENTS

The work reported herein was supported, in part, by the Explorations in Cyber International Relations (ECIR) project funded by the Office of Naval Research (ONR) contract number N00014-09-1-0597. Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the Office of Naval Research.

REFERENCES

[Camina 2010] Camina, Steven. *A Comparison of Taxonomy Generation Techniques Using Bibliometric Methods: Applied to Research Strategy Formulation*. EECS Thesis, Massachusetts Institute of Technology, 2010.

BIBLIOGRAPHY

[Blaschke 2002] Blaschke, C., Valencia, A. *Automatic Ontology Construction from the Literature*. Genome Informatics, Volume 13, 2002, pp. 201-213.

[Chuang et al. 2002] Chuang S., Chien L., *Towards Automatic Generation of Query Taxonomy: A Hierarchical Query Clustering Approach*. Academia Sinica, Taipei, 2002

[Cilibrasi & Vitanyi 2007] Cilibrasi R.L., Vitanyi P. M. B., *The google similarity distance*. IEEE Trans. Knowledge and Data Engineering Vol 19, Number 3, 2007, pp 370-383.

[Daim 2006] Daim, T.U., et al. *Forecasting emerging technologies: Use of bibliometrics and patent analysis*. Technological Forecasting and Social Change, Volume 73, Issue 8, 2006.

[Feldman 1998] Feldman, R., Dagan, I., Hirsh, H. *Mining Text Using Keyword Distributions*. Journal of Intelligent Information Systems, Volume 10, Issue 3, 1998.

- [Firat et al. 2008] Firat, A., Woon W., Madnick S. *Technological Forecasting – A Review*. Composite Information Systems Laboratory (CISL), Massachusetts Institute of Technology, 2008.
- [Glanzel 1996] Glanzel, W., *The need for standards in bibliometric research and technology*. Scientometrics, Akademiai Kiado, Volume 35, Number 2, 1996.
- [Henschel et al. 2009] Henschel A., Woon W., Wachter, T., Madnick, S. *Comparison of Generality Based Algorithm Variants for Automatic Taxonomy Generation*. Composite Information Systems Laboratory (CISL), Massachusetts Institute of Technology, 2009.
- [Heymann 2006] Heymann, P., Garcia-Molina, H., *Collaborative Creation of Communal Hierarchical Taxonomies in Social Tagging Systems*. InfoLab Technical Report, Stanford University, 2006.
- [Kostoff 2000] Kostoff, R., et al. *Fullerene Data Mining Using Bibliometrics and Database Tomography*. American Chemical Society, 2000.
- [Kostoff 2001] Kostoff R., et al. *Citation Mining: Integrating Text Mining and Bibliometrics for Research User Profiling*. Journal of the American Society for Information Science and Technology, 2001.
- [Krishnapuram 2003] Krishnapuram, R., Kimmamuru K., *Automatic Taxonomy Generation: Issues and Possibilities*. Lecture Notes in Computer Science, Springer, Berlin, 2003.
- [Martino 1993] Martino, J. . *Technological Forecasting for Decision Making, 3rd ed*. Mc-Graw-Hill, New York, 1993.
- [Martino 2003] Martino, J. *A Review of Selected Recent Advances In Technological Forecasting*. Technological Forecasting and Social Change. Vol 70, Number 8, October 2003, pp. 719-733.
- [Narin 1996] Narin, F., Hamilton, K. *Bibliometric Performance Measures*. Scientometrics. Akademiai Kiado, Volume 36 Number 3, 1996. 88
- [Okubo 1997] Okubo Y., *Bibliometric Indicators and Analysis of Research Systems: Methods and Examples*. OECD Science, Technology and Industry Working Papers, Number 1, 1997.
- [Porter 1991] Porter, A., et al. *Forecasting and Management of Technology*. Wiley-Interscience, New York, 1991.
- [Porter 2005] Porter, A., Cunningham S.. *Tech Mining*. Wiley-Interscience, New York, 2005.
- [Porter 2007] Porter, A., *How “Tech Mining” Can Enhance R&D Management*. Research Technology and Management, Mar-Apr 2007.
- [Sanchez 2004] Sanchez, D., Moreno, A., *Automatic Generation of Taxonomies from the WWW*. Practical Aspects of Knowledge Management, Volume 3336, 2004, pp 208-219.
- [Schwarzkopf et al. 2007] Schwarzkopf, E., et al. *Mining the Structure of Tag Spaces for User Modeling*. Data Mining for User Modeling, International Conference on User Modeling, Greece, 2007
- [Small 2006] Small, H., *Tracking and Predicting Growth Areas in Science*. Scientometrics, Akademiai Kiado, Hungary, 2006.
- [van Raan 1996] van Raan, A.F.J., *Advanced Bibliometric Methods as Quantitative Core of Peer Review Based Evaluation and Foresight Exercises*. Scientometrics, Elsevier Science, Oxford, 1996.
- [Verbeek 2002] Verbeek A., et al. *Measuring progress and evolution in science and technology – I: The multiple uses of bibliometric indicators*. International Journal of Management Reviews, Volume 4 Issue 2, 2002
- [Vidican et al. 2009] Vidican G., Woon, W., Madnick, S. *Measuring Innovation Using Bibliometric Techniques: The Case of Solar Photovoltaic Industry*. Composite Information Systems Laboratory (CISL), Massachusetts Institute of Technology, 2009.
- [Woon et al 2009(1)] Woon, W., Henschel, A., Madnick, S. *A Framework for Technology Forecasting and Visualization*. Composite Information Systems Laboratory (CISL), Massachusetts Institute of Technology, 2009
- [Woon & Madnick 2008] Woon, W., Madnick. S. *Asymmetric Information Distances for Automated Taxonomy Creation*. Composite Information Systems Laboratory (CISL), Massachusetts Institute of Technology, 2008.
- [Woon et al. 2009(2)] Woon W., Zeineldin, H., Madnick, S. *Bibliometric Analysis of Distributed Generations*. Composite Information Systems Laboratory (CISL), Massachusetts Institute of Technology, 2009.
- [Ziegler 2009] Ziegler, B. *Methods for Bibliometric Analysis of Research: Renewable Energy Case Study*. Composite Information Systems Laboratory (CISL), Massachusetts Institute of Technology, 2009.