Explorations in Cyber International Relations
Massachusetts Institute of Technology Harvard University

Exploring Terms and Taxonomies Relating to the Cyber International Relations Research Field: Or Are "Cyberspace" and "Cyber Space" the Same?

Steven Camiña

Department of Electrical Engineering and Computer Science
Massachusetts Institute of Technology

Stuart E. Madnick

Sloan School of Management
Massachusetts Institute of Technology

Nazli Choucri

Political Science Department
Massachusetts Institute of Technology

Wei Lee Woon

Masdar Institute of Science and Technology
Abu Dhabi, UAE

August 2011

This material is based on work supported by the U.S. Office of Naval Research, Grant No. N00014-09-1-0597. Any opinions, findings, conclusions or recommendations therein are those of the author(s) and do not necessarily reflect the views of the Office of Naval Research.

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 <u>supposed to mean</u> (somewhat complicated by the fact that they often contradict each other), they do not tell you <u>how people are actually using them</u>.

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 Village[1] is a combination of three online databases: Compendex, Inspec and NTIS. Compedex 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

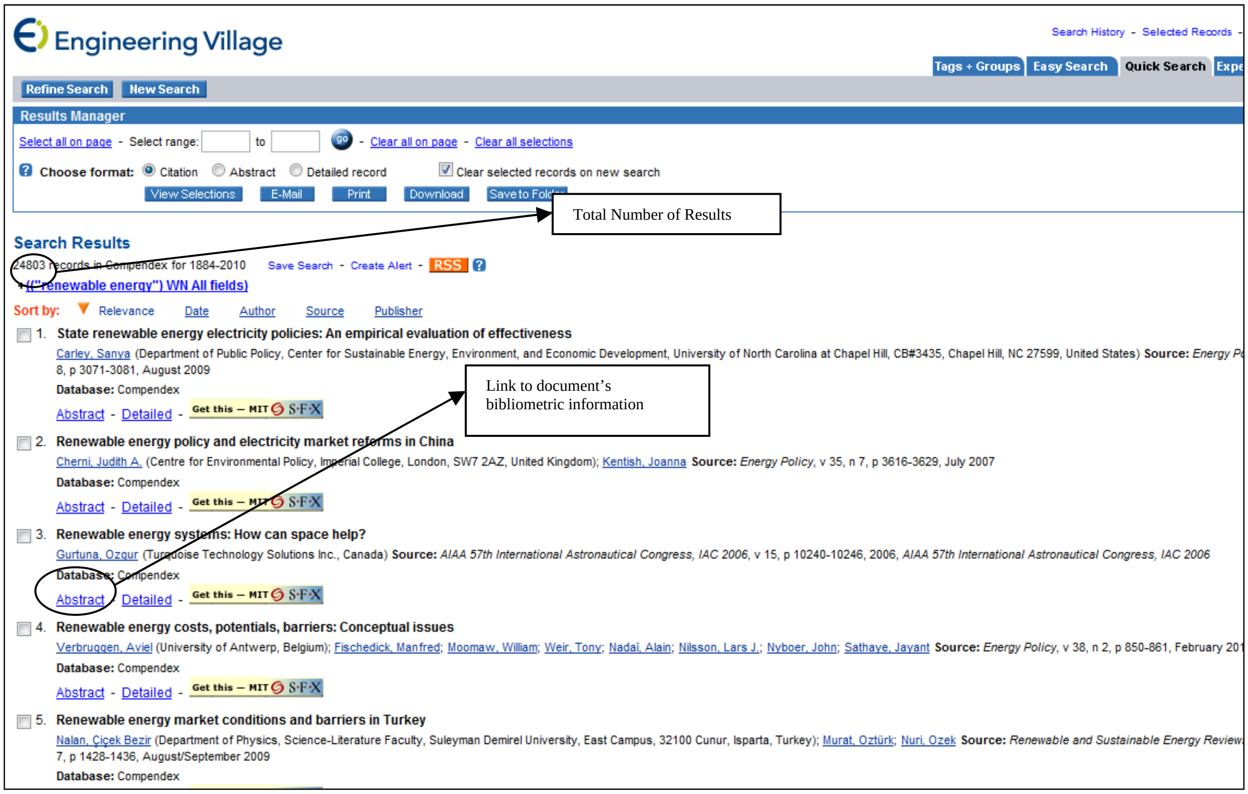

Figure 1: Screenshot of Results Page in Engineering Village. Highlighted within the figure are the locations in the website where the total number of results are shown and a link to a document's bibliometric information

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 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.

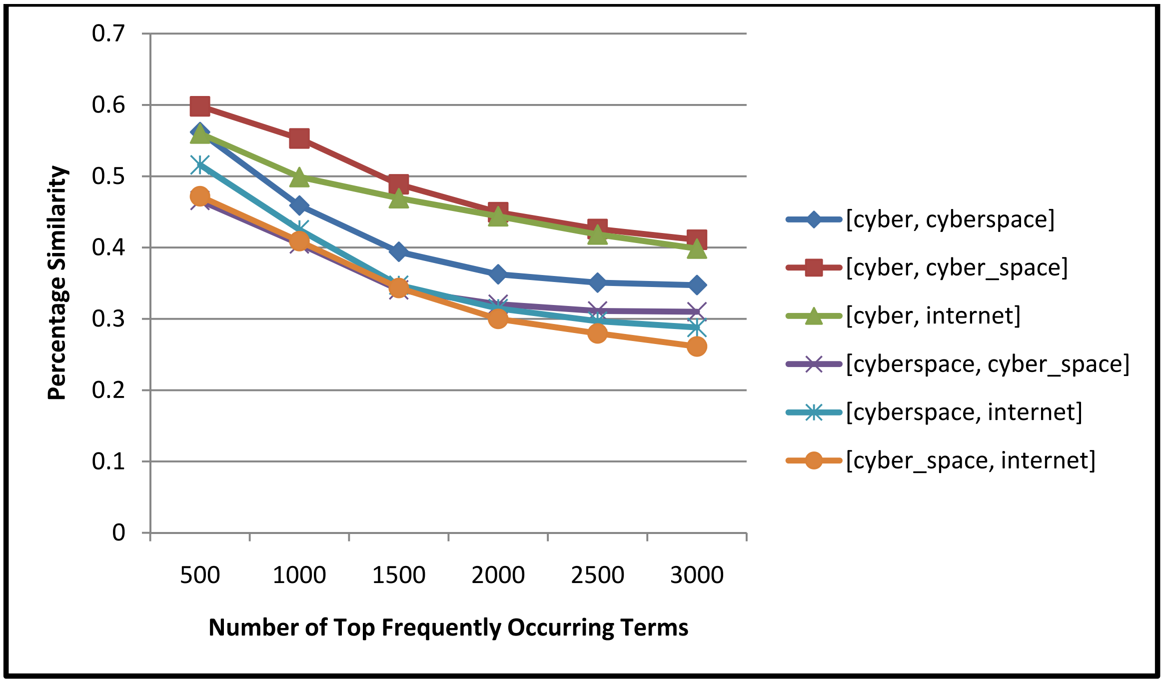

Figure 2: Percentage Similarity between terms in each of the data sets using the top 500-3000 most frequently occurring terms in the data sets

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

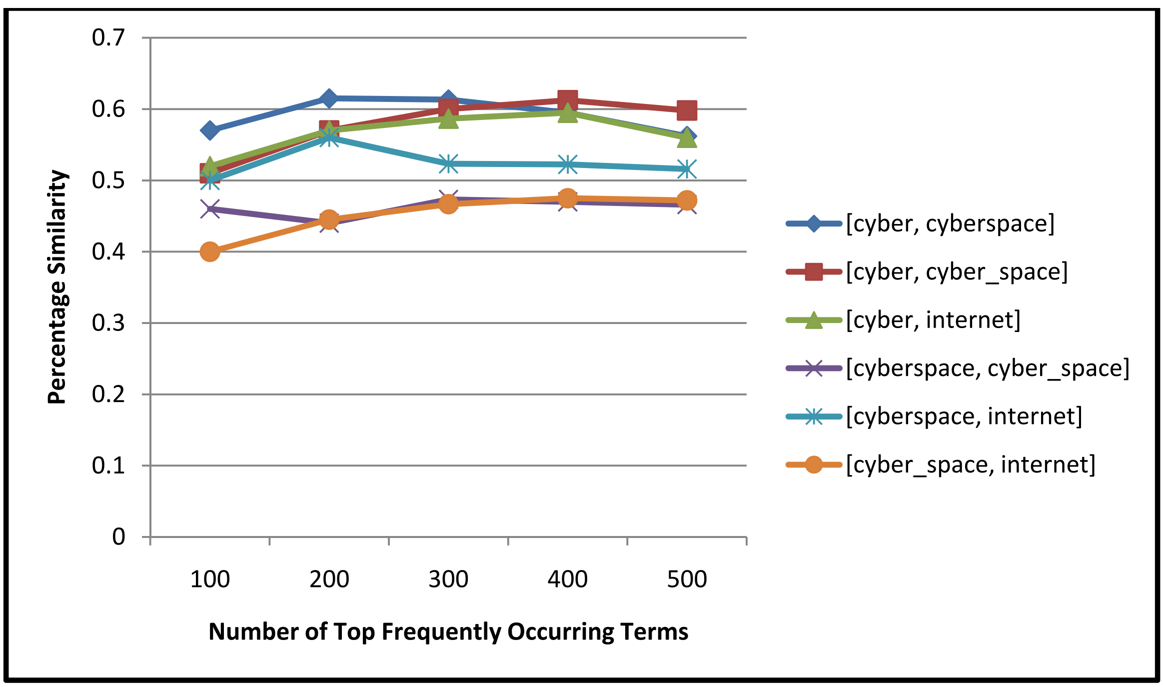

Figure 3: Percentage Similarity between terms in each of the data sets using the top 100-500 most frequently occurring terms in the data sets

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
abstracting
access control
accidents
acoustic signal processing
acoustics
ad hoc networks
adaptive algorithms
adaptive control systems
adaptive systems
administrative data processing
agents
agglomeration
alarm systems
algorithms
america
amplitude modulated
analytical models
animated movies
animation
antennas
anthropometry
applications
apriori
arsenic compounds
artificial intelligent
artificial life
audio acoustics
audio systems
audition
augmented reality
authentication
automata theory
automatic generation
automation
autonomous agents
autonomous behaviors
bandwidth
basic theory
bayesian network
behavior modeling
behavioral research
benchmarking
best efforts
biology
bipartite graphs
blind source separation
blogospheres
broadband networks
broadband services
broadcasting
calculations
calibration
cameras
campus network
cellular phones
cellular telephone systems
channel capacity
chaos theory
character recognition
charge coupled devices
china
classification (of information)
client server computer systems
closed loop control systems
cluster head
clustering
codes (standards)
codes, symbolic
coding errors
cognitive systems
collaboration systems
collaborative designs
collaborative filtering
collaborative work
collision avoidance
color
color image processing
combinatorial mathematics
command and control systems
commerce
communication
communication channels (information theory)
communication overheads
communication sessions
communications systems
competition
complex systems
computational complexity
computational geometry
computational intelligence
computational linguistics
computational methods
computational science
computer aided design
computer aided instruction
computer aided software engineering
computer animations
computer architecture
computer crime
computer forensics
computer games
computer graphics
computer hardware
computer hardware description languages
computer integrated manufacturing
computer monitors
computer music
computer networks
computer operating systems
computer privacy
computer programming languages
computer programs
computer securities
computer simulation
computer simulation languages
computer software
computer software reusability
computer supported cooperative work
computer system firewalls
computer systems
computer systems programming
computer technology
computer viruses
computer vision
computer worms
computers
computers - applications
conceptual frameworks
concurrency control
concurrent engineering
conformal mapping
congestion control (communication)
constraint theory
consumer electronics
content based retrieval
context information
context-aware
context-aware services
contracts
control
control of networks
control systems
control theory
convergence (mathematics)
convergence of numerical methods
copying
copyrights
correlation methods
cost benefit analysis
cost effectiveness
cost reduction
costs
critical component
cross-cultural study
cryptographic algorithms
cryptography
cultural difference
current technology
current trends
curricula
customer loyalty
customer satisfaction
cyber communities
cyber crimes
cyber spaces
cybercrime
cybernetics
cyberspace
data acquisition
data communication systems
data compression
data handling
data integrations
data managements
data privacy
data processing
data recording
data reduction
data securities
data sets
data storage
data storage equipment
data structures
data transfer
data visualization
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
knowledge management
knowledge representation
landforms
language processing
large scale systems
laws and legislation
learning algorithms
learning systems
legal frameworks
level of details
libraries
life-cycle
light
light measurement
linguistics
linux - operating systems
liquid crystal displays
local area networks
logic programming
low costs
low-power
machine design
machine-learning
malicious activities
malicious software
man machine systems
management
management - information systems
management systems
manipulators
maps
marketing
markov processes
mathematical models
mathematical transformations
matrix algebra
maximum likelihood estimation
medical applications
medical computing
medical imaging
medical records
message passing
metadata
microwave antennas
microwaves
military applications
military communications
military operations
mining
mobile ad hoc networks
mobile ad hoc networks (manet)
mobile agents
mobile computers
mobile devices
mobile nodes
mobile phones
mobile robots
mobile telecommunication systems
mobile users
modal analysis
models
modernization
modulation
monitoring
monitoring system
motion picture experts group (mpeg)
motion picture experts group standards
motion pictures
multi dimensional
multi-agent
multi-agent system
multicasting
multi-hop communications
multimedia services
multimedia systems
multiple cameras
multiple sources
multiplexing
museums
nanotechnologies
nash equilibrium
natural frequencies
natural resources
navigation
negative impacts
network architecture
network attacks
network intrusion detections
network management
network monitoring
network operators
network protocols
network resources
network size
network technologies
network topology
network traffic
networked systems
networks security
neural networks
new approaches
new concept
next generation networks
normal-hearing (nh)
novel methods
numerical methods
object oriented programming
object recognition
ocean engineering
oceanography
offline
online communities
online conferencing
online discussions
online forum
on-line gamings
online learning
online shopping
online systems
ontology
open source software
open sources
open systems
operational modeling
operational systems
optical communication
optical data processing
optical fibers
optical systems
optimization
outsourcing
p2p system
packet networks
paradigm shifts
parallel processing systems
parameterization
patient monitoring
pattern matching
peer-to-peer networks
performance
personal computers
personal digital assistants
personal information
personalization
personnel
personnel training
pervasive computing
petri nets
philosophical aspects
photography
photons
physical world
pixels
plain text
planning
policy-makers
polynomials
portable equipment
portals
printing
printing presses
probability
probability density function
probability distributions
problem oriented languages
problem solving
process control
process information
product design
production control
productivity
profile
program processors
programmable logic controllers
project management
proof of concepts
protocol designs
prototype implementation
prototype system
public key cryptography
public policy
public space
qos requirements
quality assurance
quality control
quality of service
quantitative method
query languages
radio broadcasting
radio communication
random processes
real time systems
real times
real-space
real-world
reasoning process
recommendation systems
redundancy
regression analysis
regulatory compliance
reliability analysis
reliable
remote control
remote education
remote sensing
remote users
research
research activities
research and development management
research communities
research results
residual energy
resource allocation
resource sharing
response time (computer systems)
retransmissions
reusability
revenue
rfid technology
risk analysis
risk assessment
risk management
road maps
roads and streets
robotics
role-playing games
rom
rough set theory
routers
routing protocols
sales
satellite communication systems
scada systems
scheduling
school buildings
search engines
security infrastructures
security levels
security management
security mechanisms
security of data
security protocols
security requirements
security services
security situation
security systems
self-organize
semantic information
semantic web
semantics
semiconductor quantum dots
sensor data fusion
sensor fusion
sensor networks
sensor nodes
sensors
sensory perception
servers
service discovery
service provider
service quality
service users
set theory
sign language
signal detection

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

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

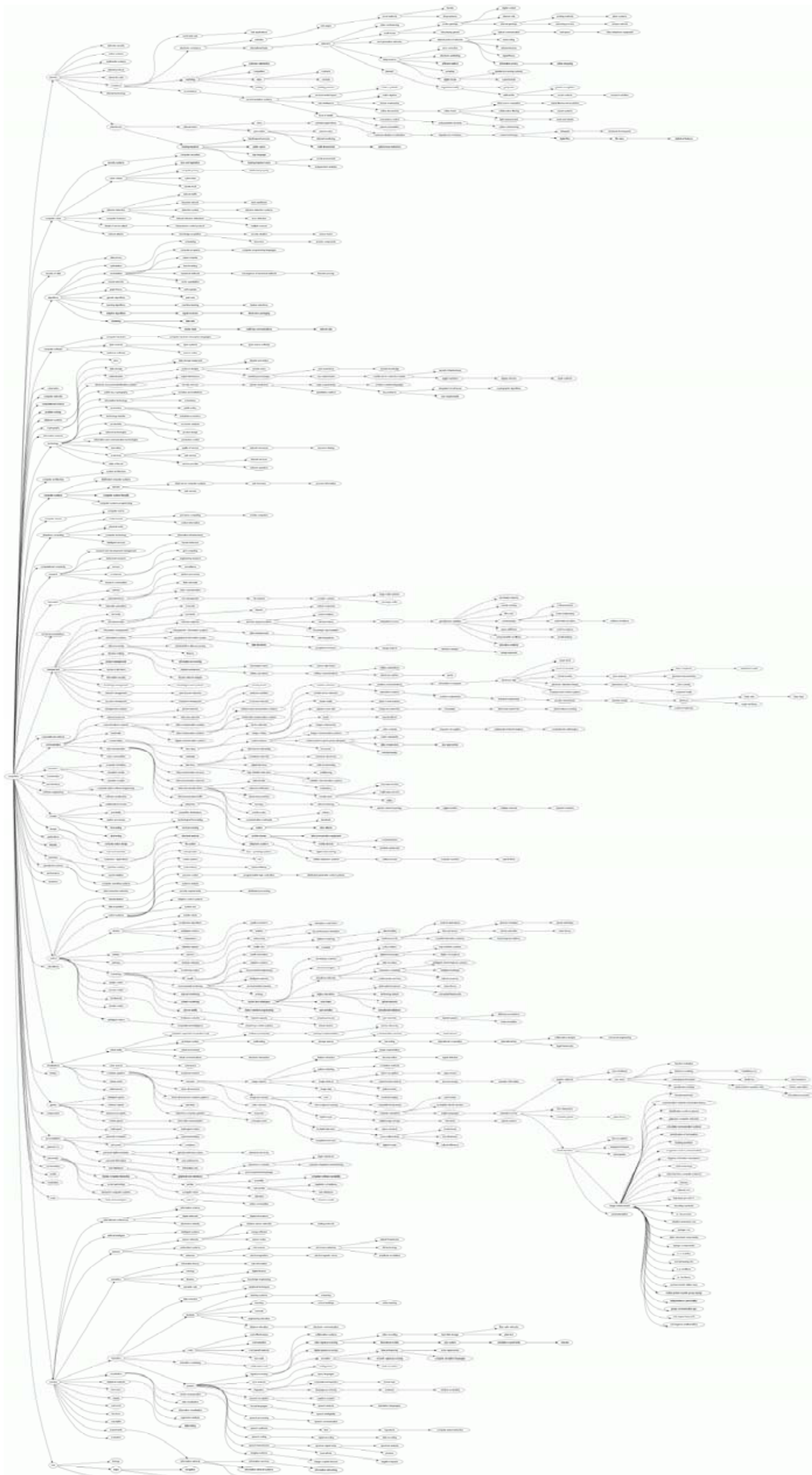

Figure 4: "Cyber" H-CC Taxonomy

Figure 5: “Cyberspace” H-CC Taxonomy

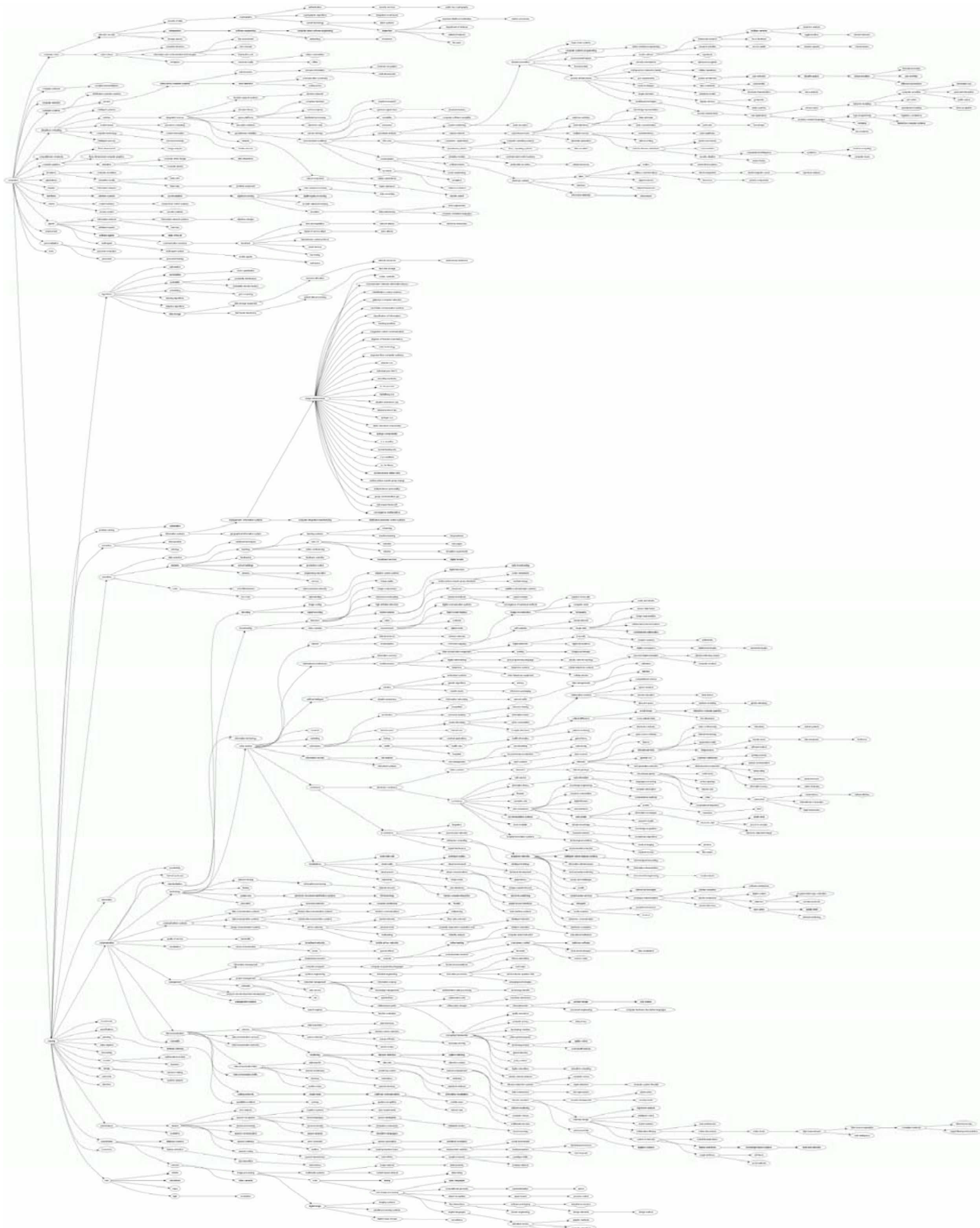

Figure 6: “Cyber space” H-CC Taxonomy

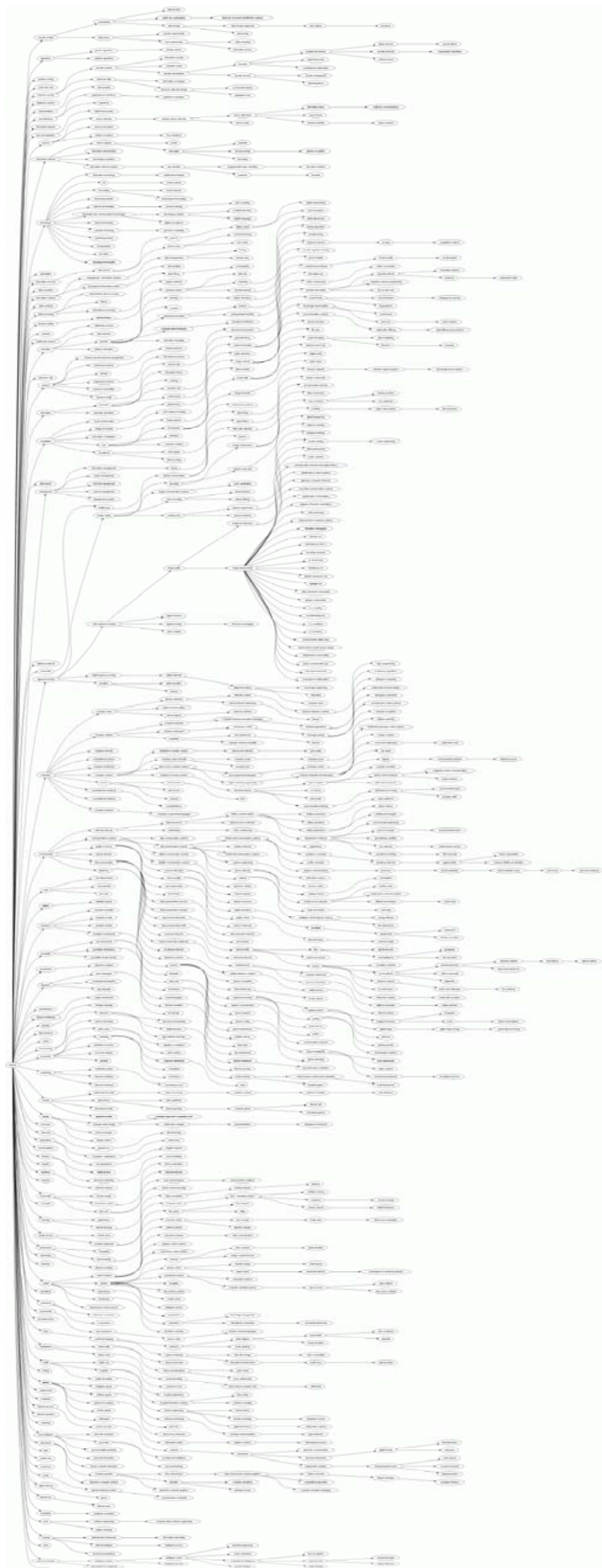

Figure 7: “Internet” H-CC Taxonomy

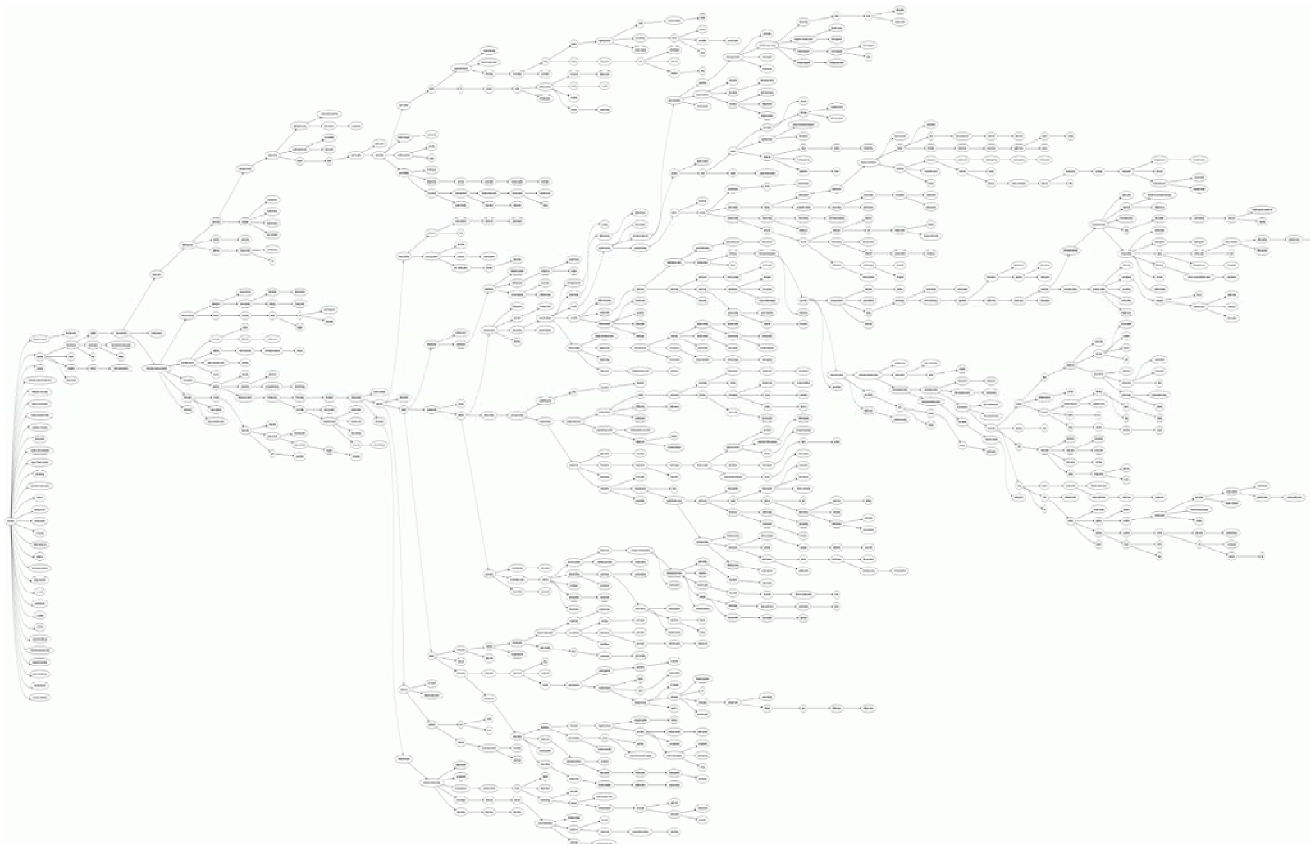

Figure 8: "Cyber" D-SC Taxonomy

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 at 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 Possiblities.* 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.