MIT ESD

Massachusetts Institute of Technology
Engineering Systems Division

ESD Working Paper Series

Towards better understanding Cybersecurity: Or are "Cyberspace" and "Cyber Space" the same?

Stuart Madnick
John Norris Maguire Professor of Information Technology and Professor of Engineering Systems
MIT Sloan School of Management and MIT School of Engineering
Massachusetts Institute of Technology
Email: smadnick@mit.edu

Nazli Choucri
Professor of Political Science and Associate Director of the Technology & Development Program
MIT Department of Political Science
Massachusetts Institute of Technology
Email: nchoucri@mit.edu

Steven Camiña
Research Assistant
MIT Department of Electrical Engineering and Computer Science
Massachusetts Institute of Technology
Email: stevenlc@gmail.com

Wei Lee Woon
Associate Professor
Electrical Engineering and Computer Science
Masdar Institute of Science and Technology
Email: wwoon@masdar.ac.ae

ESD-WP-2014-28 October 2014 esd.mit.edu/wps

Towards better understanding Cybersecurity: or are "Cyberspace" and "Cyber Space" the same?

Stuart Madnick
Nazli Choucri
Steven Camiña
Wei Lee Woon

Working Paper CISL# 2012-09

November 2012

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

Towards better understanding Cybersecurity: or are "Cyberspace" and "Cyber Space" the same?

Stuart Madnick
<smadnick@mit.edu>
MIT Sloan School of Management
and MIT School of Engineering
Nazli Choucri
<nchoucri@MIT.EDU>
MIT Dept of Political Science
Steven Camiña
<stevenlc@gmail.com>
MIT Dept of Electrical Engineering and Computer Science
Wei Lee Woon
<wlwoon@gmail.com>
Masdar Institute of Science and Technology

ABSTRACT

Although there are many technology challenges and approaches to attaining cybersecurity, human actions (or inactions) also often pose large risks. There are many reasons, but one problem is whether we all "see the world" the same way. That is, what does "cybersecurity" actually mean – as well as the many related concepts, such as "cyberthreat," "cybercrime," etc. 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. If we are to have an effective solution, it is important that all the parties understand each other – or, at least, understand that there are different perspectives.

For the purpose of this paper and to demonstrate our methodology, we consider the case of the words, "cyberspace" and "cyber space." 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 usage of these words (as determined by the taxonomies that would be constructed by our algorithms) 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 cybersecurity is so new, understanding the field and how people think about it (as evidenced by their actual usage of terminology, and how usage changes over time) is an important goal. Our approach helps to illuminate these understandings.

1. INTRODUCTION

Although there are many technology challenges and approaches to attaining cybersecurity, ihuman actions (or inactions) also often pose large risks. There are many reasons, but one problem is whether we all "see the world" the same way. That is, what does "cybersecurity" actually mean – as well as the many related concepts, such as "cyberthreat," "cybercrime," etc. 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. . If we are to have an effective solution, it is important that all the parties understand each other – or, at least, understand that there are different perspectives.

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 credible academic online publication databases. In particular, this paper analyzes the online publication databases within Engineering Village, namely Compendex and Inspec.

For the purpose of this paper and to demonstrate our methodology, we consider the words, "cyberspace" and "cyber space" (the *seed terms* used.) 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 usage of these words (as determined by the taxonomies that would be constructed by our algorithms) would be basically the same. As it turned out, they were quite different, both in overall shape and groupings within the taxonomy

2. SOURCES OF DATA

2.1 Sources Used

Engineering Village[1] is a combination of several online publication databases, in particular Compendex and Inspec. Compendex is a comprehensive bibliographic database of scientific and technical engineering research, covering all engineering disciplines. Compendex includes over 5 million summaries of journal articles and conference proceedings and there are 220,000 new additions every year.

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.

2.2 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

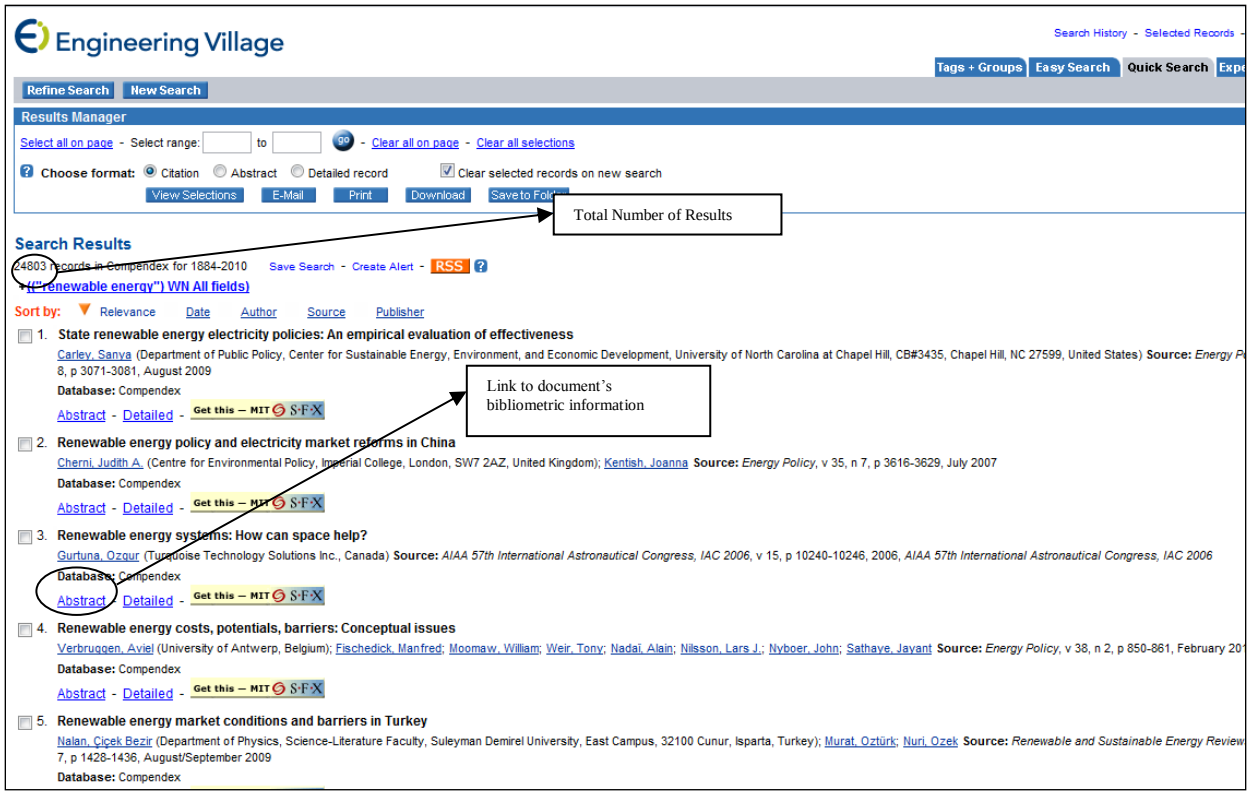

Figure 1: Screenshot of Results Page in Engineering Village. Highlighted within the figure are the locations in the website where the total number of results are shown and a link to a document's bibliometric information

Software has been developed to extract (often referred to as "scraping") each document's bibliometric information from the website. Specifically, we use the document's title, abstract, and keywords. Each document has multiple controlled and uncontrolled keywords, which we refer to as the *terms* of each document.

The bibliometric information of the various articles scraped from online publication databases is then stored into a local file, which can then manipulated. 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 further 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. Use the terms and data set to generate a taxonomy, which is a hierarchical organization of the terms.

In gathering the results using the seed terms mentioned previously, both Compendex and 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 we chose to gather bibliometric information from.

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

Seed Term	Compendex Document Count	Inspec Document Count
"cyberspace"	**983**	637
"cyber space"	**968**	720

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 "cyberspace" and "cyber space".

2.4 Terms in Each Data Set and Terms in Common

The next step is to gather all the keywords related to the *seed term*, which in this case was either "cyberspace" or "cyber space." This is done by gathering all the keywords from the documents related to the seed term. Of course, duplicate occurrences of keywords are removed. 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
"cyberspace"	3,488
"cyber space"	4,717

Table 2: Summary of Terms Contained in Each Data Set

Of these term sets, 886 terms were found to be common across both data sets. These terms represent the concepts common to both the "cyberspace" and "cyber space" landscapes, which we used for further analysis.

3. TAXONOMY GENERATION

3.1 Algorithms used for Taxonomy Generation

The next step was to generate both taxonomies and compare the results, using the 886 terms in common mentioned above. The algorithms used for taxonomy generation were motivated by and described in detail in [Camina 2010]. These algorithms make use of the Heymann algorithm, closeness centrality, cosine similarity metric (which we refer to as H-CC.) [Heymann 2006] [Sanchez 2004] [Woon et al 2009] [Ziegler 2009] Without going into the details, the taxonomy generation process involves the following steps:

1. Creating a co-occurrence matrix with the count of the number of times that the two terms occurred in the same document – which is a measure of the degree of "closeness" of the two terms.
2. Determining the "root" of the taxonomy by finding which term is most central in the graph represented by the co-occurrence matrix.
3. Filling out the first level of the taxonomy by finding the terms "closest" to the root term, then the second level is filled out by finding the terms "closest" to each of the terms in the first level, etc.

The two resulting taxonomies, for "cyberspace" and "cyber space," are shown in Figure 2a and 2b below. It is not expected that the reader will be able to see each of the 886 terms in each of these taxonomies. The point is to note that the shapes of these taxonomies are quite different. The key differences will be discussed in the following sections.

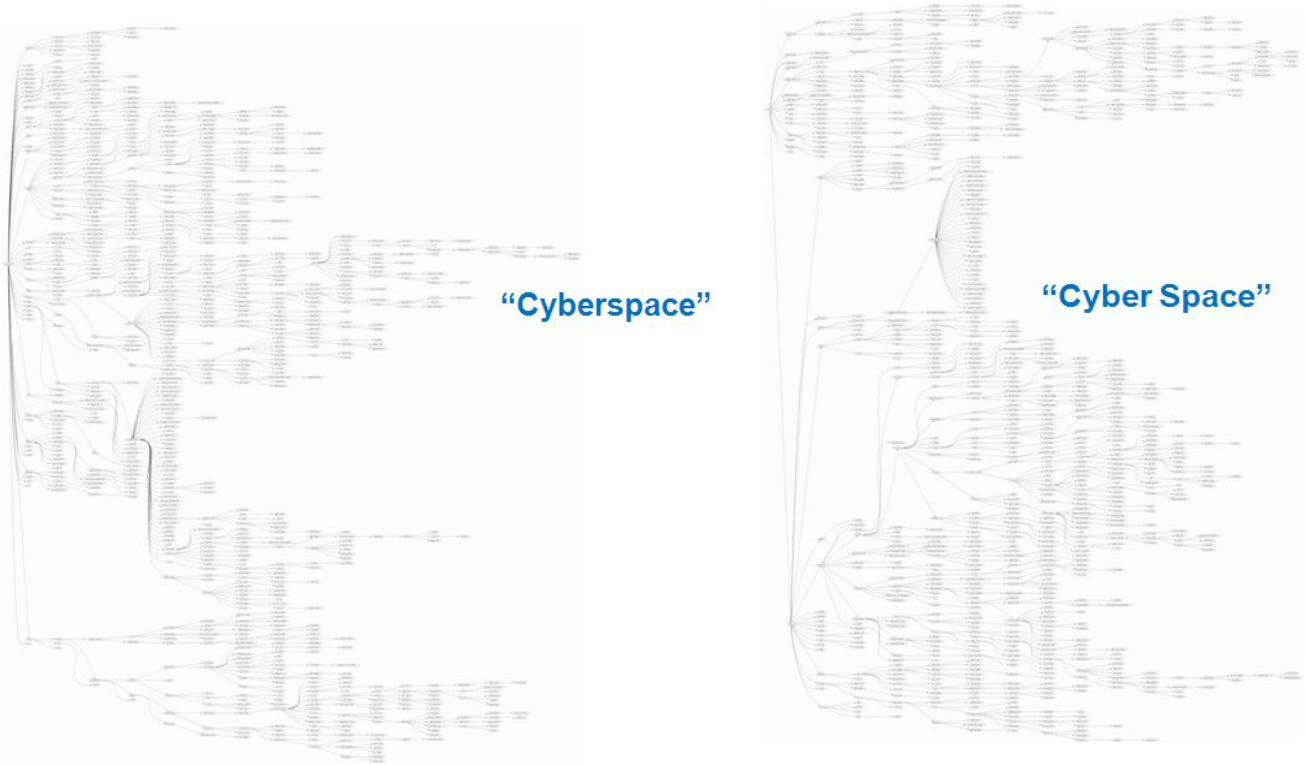

Figure 2: Taxonomies for (a) "Cyberspace" and (b) "Cyber Space"[2]

3.2 Root Terms in Taxonomies Generated

In our taxonomy generation, the *seed term* used to generate the data set is not necessarily the same as the *root term*, or term at the root 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.

Table 3 summarizes the root term found for each taxonomy generated.

Seed Term Used to Generate Data Set	Taxonomy Root Term
"cyberspace"	Cyberspace
"cyber space"	Computers

Table 3: Root Terms for Each Taxonomy Generated

3.3 Comparison of Taxonomies Generated

Table 4 below shows pairwise comparisons between each of the two taxonomies generated. 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 mentioned 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.

[2] High resolution copies of the GIF files for Figure 2a and 2b 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).

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
"cyberspace"	"cyber space"	19.41%

Table 4: Percentage Similarity of Taxonomies Generated using H-CC algorithm

Based on our prior expectations (i.e., that "cyberspace" and "cyber space" were basically the same), it was surprising how low was the percent of similar links, only about 19%. But this was not surprising given the significant differences in shape evident in Figures 2a and 2b.

3.4 Analysis of Taxonomy Differences

3.4.1 "Cyberspace" Taxonomy

As mentioned above, Figure 2a shows a birds-eye view of the "cyber space" taxonomy. Some of the interesting observations and distinctions about the taxonomy are:

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. There is a cluster with "internet" as the root

3.4.2 "Cyber space" Taxonomy

Figure 2b shows a birds-eye view of the "cyber space" taxonomy generated. Some of the interesting observations about the taxonomy are:

1. The root term of the taxonomy is "computers"
2. This taxonomy included "telecommunication", "speech" and "algorithms" clusters
3. There is a cluster with "technology" as the root, leading to terms such as "information technology", "cyberspaces", and "innovation"
4. There is a cluster with "disaster prevention" as the root, leading to terms such as "environmental impact", and "security infrastructure"
5. There also appears to be a lot of noise / nonsense links in this taxonomy. For example, 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.

4. CONCLUSIONS AND FUTURE RESEARCH

4.1 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 the taxonomies generated and displayed in Figures 2a and 2b are clearly very different.

What is the reason (or reasons)? Some hypotheses might be: (a) authors of papers in different academic fields use different words (e.g., policy people vs. technology people), (b) authors from different parts of the world use different words, (c) the words were used (with different meanings) in different time periods, etc. – the reader may have other hypotheses to suggest. Through further investigation, we were able to answer this question, but due to lack of space, and to leave a bit of anticipation for the reader, this will not be included in this paper.

4.2 Future Research

The terms "cyberspace" and "cyber space" were used just to demonstrate the process … and because they led to some interesting results. We plan to use these techniques to study other terms, such as "cybersecurity," "cyberthreat," "cybercrime," etc.

In addition, some areas for future extension of the methodology include (with very abbreviated explanations):

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

4.2.2. Choice of type of sources: We used a database of academic publications. We could use blogs and news which could provide much more timely information. What would that look like?

4.2.3. Choice of language: We have mainly focused on English publications, what if we included publications from other languages - possibly translating the keywords into English.

4.2.4. 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?

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

4.2.6. Algorithms: We have experimented with various algorithms for the automated generation of taxonomies. The H-CC combination was used in this paper. How would the results differ if other possible algorithms were used?

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

4.2.8. "Face validity": It would be good to show our automatically generated taxonomies to more Subject Matter Experts (SMEs) to get their reactions.

ACKNOWLEDGEMENTS

This work is funded by the Office of Naval Research under award number N00014-09-1-0597. Any opinion or findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the Office of Naval Research

REFERENCES

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

[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.

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

[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.

[Krishnapuram 2003] Krishnapuram, R., Kimmamuru K., *Automatic Taxonomy Generation: Issues and Possiblities.* Lecture Notes in Computer Science, Springer, Berlin, 2003.

[Sanchez 2004] Sanchez, D., Moreno, A., *Automatic Generation of Taxonomies from the WWW.* Practical Aspects of Knowledge Management, Volume 3336, 2004, pp 208-219.

[Woon et al 2009] Woon, W., Henschel, A., Madnick, S. *A Framework for Technology Forecasting and Visualization.* 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.