16th ICCRTS

"Collective C2 in Multinational Civil-Military Operations"

Title

Lexical Link Analysis for the Haiti earthquake Relief Operation Using Open Data Sources

Topics

Primary: Topic 3: Information and Knowledge Exploration

Alternatives: Topic 5: Collaboration, Shared Awareness, and Decision Making

Topic 2: Approaches and Organizations

Authors

Dr. Ying Zhao, Dr. Shelley P. Gallup and Dr. Douglas J. MacKinnon

Distributed Information Systems Experimentation (DISE) Group

Naval Postgraduate School, Monterey, CA 93943

yzhao@nps.edu, spgallup@nps.edu, djmackin@nps.edu,

Point of Contact

Dr. Ying Zhao

Information Sciences Department
Naval Postgraduate School

1411 Cunningham Road, Root Hall 229
Monterey, CA 93943-5000
(831)-656-3789 (office)

yzhao@nps.edu

Lexical Link Analysis for the Haiti Earthquake Relief Operation Using Open Data Sources

Dr. Ying Zhao, Dr. Shelley P. Gallup, and Dr. Douglas J. MacKinnon
Distributed Information Systems Experimentation (DISE) Group
Naval Postgraduate School, Monterey, CA 93943
yzhao@nps.edu, spgallup@nps.edu, djmackin@nps.edu

1 Abstract

In the wake of the Haiti earthquake disaster, civil and military organizations engaged in vigorous relief operations to achieve rapid deployment of logistics, transport, security and medical supplies. Organizations involved in the operation collaborated with the expectation of effectiveness. To know that operations were effective - or understanding how to improve their effectiveness - would require sifting through large volumes of communications to understand how various organizations collaborated and to improve future operations.

In this paper, we will report how to leverage "Lexical Link Analysis" (LLA), as one means to achieve "System Self-awareness" (SSA), and develop this as a durable methodology. We consider that the cognitive interface between decision makers and a complex system, (e.g. the organizational behaviors in an interagency operation) may be expressed through lexical link features embedded in documents and communications. The LLA method is composed of lexical analysis and link analysis. Lexical analysis is a form of text mining in which word pairs are extracted from document sets. Link analysis is a network analysis that explores key relationships among objects. We will report how to combine both the LLA and SSA methods to sift through real-life, open source data, applied in this example to Haiti relief operations.

2 Introduction

As a matter of policy the United States mobilizes diverse responses in aid,both internally and externally (to other countries), following natural disasters. In the aftermath of the Haiti earthquake, US military and civil organizations provided rapid and extensive relief operations. Organizations involved in the operation collaborated with each other to negotiate roles, define limits, set business rules etc., in order to maximize efficiency and decrease response times. To determine that the expectations for effective and efficient operations had been attained would require in-depth analysis of collaboration, dialog and resultant outcomes. Such analysis by manual means would require long and arduous effort. We propose automation of a portion of the analytic work, asking the questions:

1. What were the roles and relationships of these organizations?
2. How the operations were actually conducted?

For the lead military organization, there is an assumption that because the operational commander had resources and trained personnel, the expected outcome would be timely deployment of logistics, transport, security and medicine. Also, there is an assumption among participating military organizations that due to their similar training and skills and shared culture, their collaboration

would result in higher effectiveness than civilian organizations. While anecdotal evidence supports this view, proving it as a fact is very difficult without tools such as LLA.

The challenge is to sift through the data that are collected in real-life events to create an overall picture of how various organizations (military and civil) actually collaborated, how their interactions were developed, and how their synergies were achieved. For example, how are the findings from real-life data different from participant expectations? How can these findings be used to improve the future operations?

In this white paper, we leverage "Lexical Link Analysis" (LLA) and a related concept, "System Self-awareness" (SSA). We use data samples and analysis from open sources of Haiti operations to illustrate the method.

3 Approaches: System Self-Awareness (SSA) and Lexical Link Analysis (LLA)

We consider that the cognitive interface between decision makers and a complex system, e.g. an organizational behavior in an interagency operation may be described in a range of terms or "features," i.e. specific vocabulary or lexicon to describe attributes and the surrounding environment and its missions and tasks that are embedded in the documents and communications.

We borrow from notions of "awareness" and implement system *self-awareness* (SSA, Gallup, et al., 2009) of a complex system as the collective and integrated understanding of system features. A related term, "situational awareness" is used in military operations and carries with it a sense of immediacy and cognitive understanding of the warfighting situation.

Lexical Analysis (LA, 2010) is a form of text mining to analyze the large numbers of features or lexicon of a complex system. Lexical Analysis (LA) can also be used in a learning mode, where meaning is constantly being "learned," updated and improved as more data become available.

Link analysis, a subset of network analysis that explores associations between objects, provides the crucial relationships between objects when collected data may not be complete.

Lexical Link Analysis (LLA, Zhao, Gallup & MacKinnon, 2010) is an extended lexical analysis that, when combined with link analysis, is capable of learning and data mining which can be used to dynamically identify and assess, as well as predict trends, patterns, and features that help predict and change future behavior.

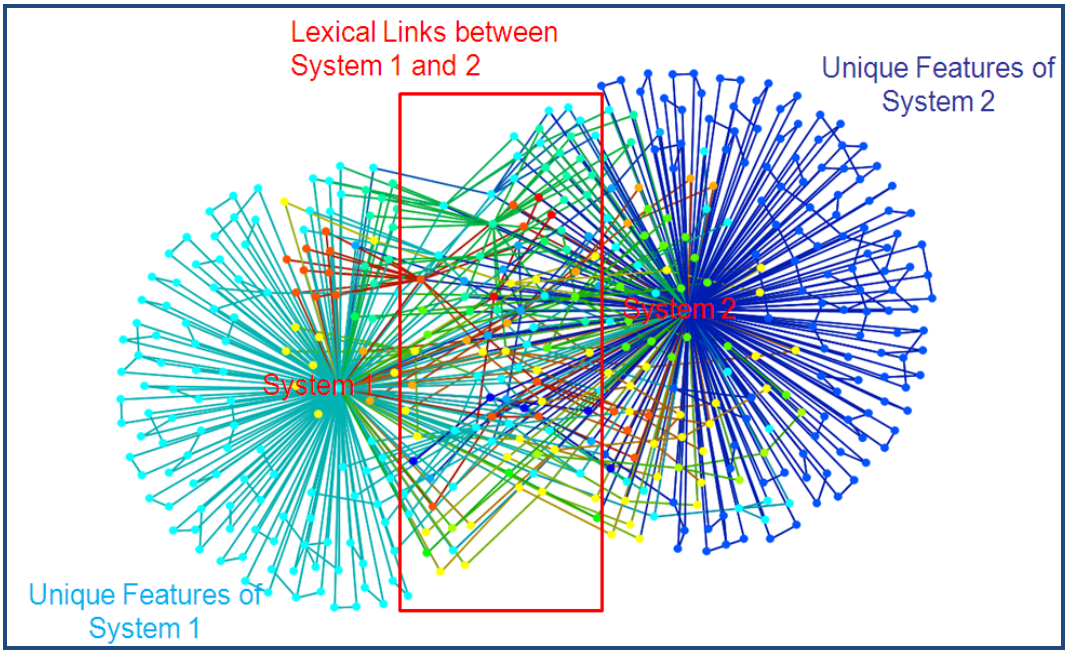

Figure 1: Lexical Links between two systems are shown in the red box.

Figure 1 shows a visualization of lexical links for two systems System 1 and System 2. Each node is a feature or word hub, each color refers to the collection of lexicon (features) to describe a concept or a topic. The overlapping area nodes refer to *lexical links* between systems. The nodes and links pointed outward, away from the other system, represent the unique features related to its system.

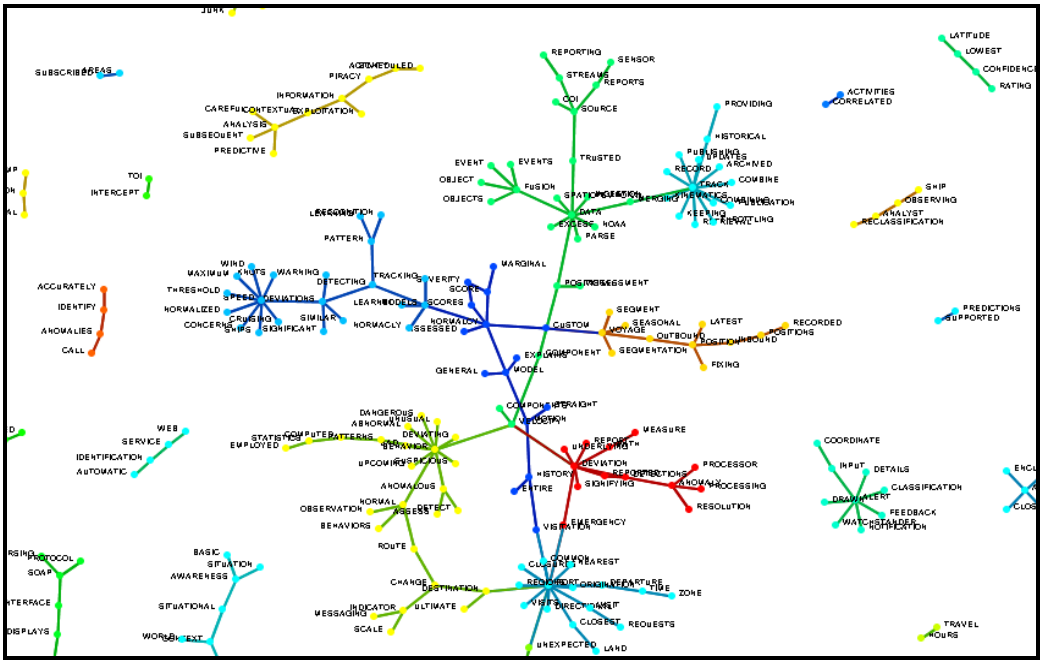

Figure 2(a): "Features" are shown as words that are linked semantically, representing word networks discovered from conceptually interrelated data sources.

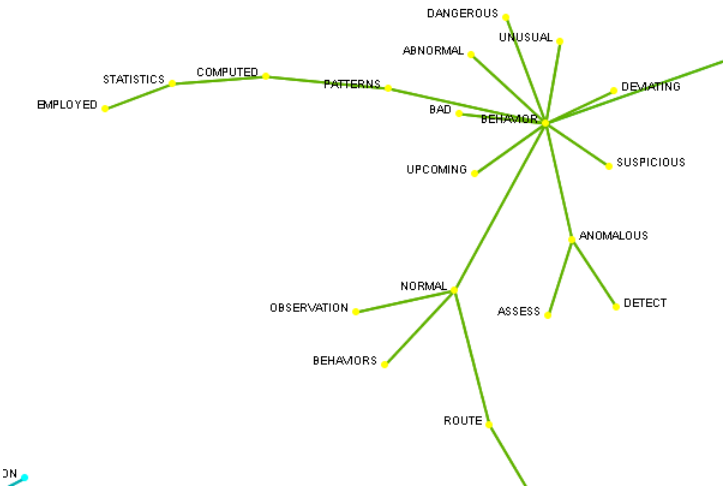

Figure 2(b): (detail of Figure 2(a): A single word hub showing linkage among dangerous, abnormal and unusual behavior.

Figure 2(a) shows a visualization of LLA (Lexical Link Analysis) with connected keywords or concepts extracted from the documents of Maritime Domain Awareness (MDA) technologies. Words are linked as word pairs that appear next to each other in the original documents. Different colors indicate different clusters of centralization among word groups. They are produced using a link analysis method - a social network grouping method (Girvan, et al., 2001), where words are connected as shown in a single color as if they are in a social community. A "hub" is formed around a word *centered* or connected with a list of other words ("fan-out" words) centered on other hub words. For instance, in Figure 2(b), the word "behavior" is centered with "suspicious, bad, dangerous, abnormal, unusual, and anomalous" etc., showing the ways to describe "behavior" in the MDA area.

In the past year, we began using LLA at the Naval Postgraduate School (NPS) in conjunction with Collaborative Learning Agents (CLA) (QI, 2009) and expanded to other tools such AutoMap (AutoMap, 2009) for improved visualizations. Results from these efforts arose from leveraging intelligent agent technology via an educational license with Quantum Intelligence, Inc. CLA is a computer-based learning agent or agent collaboration, capable of ingesting and processing unstructured data sources such as text documents and communications.

This approach is related to a number of extant tools for text mining including Latent Semantic Analysis (LSA) (Dumais, Furnas, Landauer, Deerwester &Harshman, 1988), key word analysis and tagging technology (Foltz, 2002), and intelligence analysis ontology for cognitive assistants (Tecuci, Boicu, Marcu, Barbulescu, Ayers & Cammons, 2007). What results from this process is a learning model - like an ethnographic *code book* (Schensul, Schensul & LeCompte, 1999). The topic extraction method embedded in LLA is also related to recent development of Latent Dirichlet Allocation (LDA, Blei, Ng &Jordan, 2003) where topics are also discovered but in a form of a "bags of words." Our method reveals a collection of lexical terms connected to each other semantically. What is unique here is that LLA constructs these linkages via intelligent agent technology using social network grouping methods.

In our methodology, we first applied LLA to extract relevant features and lexical links from text mining open source text documents that were published from the organizations that were involved in the Haiti operations. Combining text mining and social network analysis in LLA allowed grouping of a large number of features into concepts, topics and themes using social network community finding algorithms. The concentration of features is measured by various social network centrality measures, for example, total degree measures how many total links from a node (Wasserman & Faust, 1994). These measures were developed into metrics that can be used for evaluating interagency collaborations.

When compared with WordNet (2010) developed at Princeton University, a lexical database of English terms and their relationships is able to generate a *map* for a word that semantically connects to other words, LLA is data-driven and feature relations are discovered automatically and dynamically from the data with little use of manual extraction to go through the hefty data sets. WordNet was created manually, therefore not updated frequently and cannot be specific to a particular application domain.

3.1 Research Questions

In the long term, we are interested in applying LLA and SSA as a monitor to visualize organizational behavior in real-time; for example, examining situational awareness of interagency operations in a disaster or relief operation. What follows is a list of research questions we are interested in pursuing using these methodologies:

1) How can synergy, efficiency, and competency of an interagency operation be measured that involves many organizations with respect to military or civilian interests?
2) Can lexical links be used as metrics to measure the synergy, efficiency, competency of an interagency operation?
3) How can we extract social networks such as people, locations, and organizations from social media or social network data?
4) Do social media, such as a discussion forum or social networking tools (e.g. Facebook and Twitter) help in an interagency operation?
5) How is synergy between organizations developed over time?

3.2 An Open Source View

This paper shows a case study on how LLA was applied to samples of open source documents that were collected for the Haiti earthquake relief operation. We report here the initial study and results using LLA to address the questions in the previous section.

3.2.1 Data Collection Process

We collected ~2600 open source web pages using a cluster of High Performance Computer (HPC) nodes. The data were the news feed from 1/13 – 2/23/2010, when the Haiti earthquake and international relief operations took place. The data was sorted by timestamps, domains and organizations. The steps of data collection are summarized as follows:

- **Step 1**: Start with a list of web pages below, ask the crawler to go two levels deep to obtain all the links in the pages
  - http://twitter.com/southcomwatch
  - http://www.southcom.mil/AppsSC/factFiles.php?id=138
  - http://twitter.com/USAID_Haiti
  - http://www.inrelief.org/
- **Step 2**: Go through all the pages collected from Step 1 and sort them according to timestamps, domains and organizations.

We selected Twitter as a starting point since various organizations such as SOUTHCOM and USAID used it to handle the situations that required real-time information gathering and dissemination such in the Haiti relief efforts. Inrelief.org was a new site hosted by NPS that was of interest to us as how it worked in a real-life event.

3.2.2 Research

Our goal was to show initially that lexical links can be utilities and measures for trends of interagency operations. We collected the following lexical link measures

- Number of nodes or word hubs/features
- Number of links between the word hubs
- Number of domains or organizations
- Number of cross-domains, for example, twitter/SOUTHCOM is a cross-domain when SOUTHCOM used Twitter to communicate
- Synergy index: Defined as the number of word hubs between two organizations divided by the total number of word hubs from the two organizations.

Synergy is normally a measure obtained using the traditional quadratic assignment procedure correlation (QAP, Hubert & Schultz, 1976) used in social sciences and social network analysis. However, synergy defined here is calculated from lexical links instead of social network links which are commonly used. QAP calculates Pearson's correlation coefficient between corresponding cells of the two data matrices. QAP also tests if the association between two networks is statistically significant.

3.2.3 Example Views

In this section we show a few example visualizations that were produced in LLA. We used the tool Organization Risk Analyzer (ORA) 2.0.8 within AutoMap to produce the visualizations.

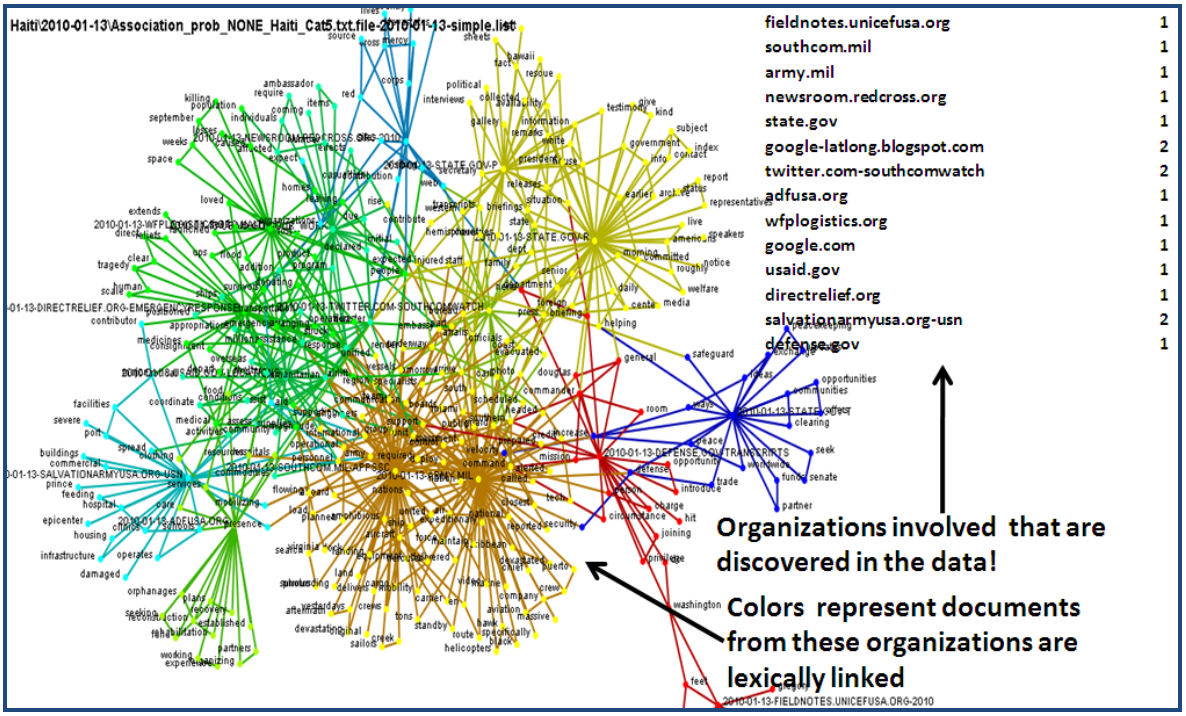

Figure 3: Synergy was low on 2010-01-13.

Figure 3 shows the lexical links on the first day of the operation. There were 14 organizations discovered for the day, each was centered on word pairs or lexical links discovered in the documents published by the organization. Some organizations were cross-domain, for example, twitter.com-southcomwatch. The colors, generated according to the social networking community finding algorithm over the word networks found in the documents, indicate word pairs with the same colors are lexically linked. In Figure 3, same-color links tend to belong to the same organizations. However, there are cases where links from two organizations share the same color which provides evidence that synergy exists between the two.

Figure 4(a) shows a closer look of Figure 3 where we found southcom.mil and army.mil were grouped into the same color because their contents were linked through the common terms including “assessment”, “deploy” and “air force”, etc. An enlarged graph in Figure 4(b) shows the lexical links between two organizations of southcom.mil and army.mil in detail.

Figure 5(a) also shows a closer look of Figure 3 where the synergies among various NGOs such as USAID, Direct Relief and WFP Logistics were also high because that their lexical links were grouped into the same color according to the common concerns of “food”, “medical”, “shelter”, “transportation” and “flood.” Figure 5(b) shows an enlarged portion of the lexical links between Direct Relief and WFP Logistics.

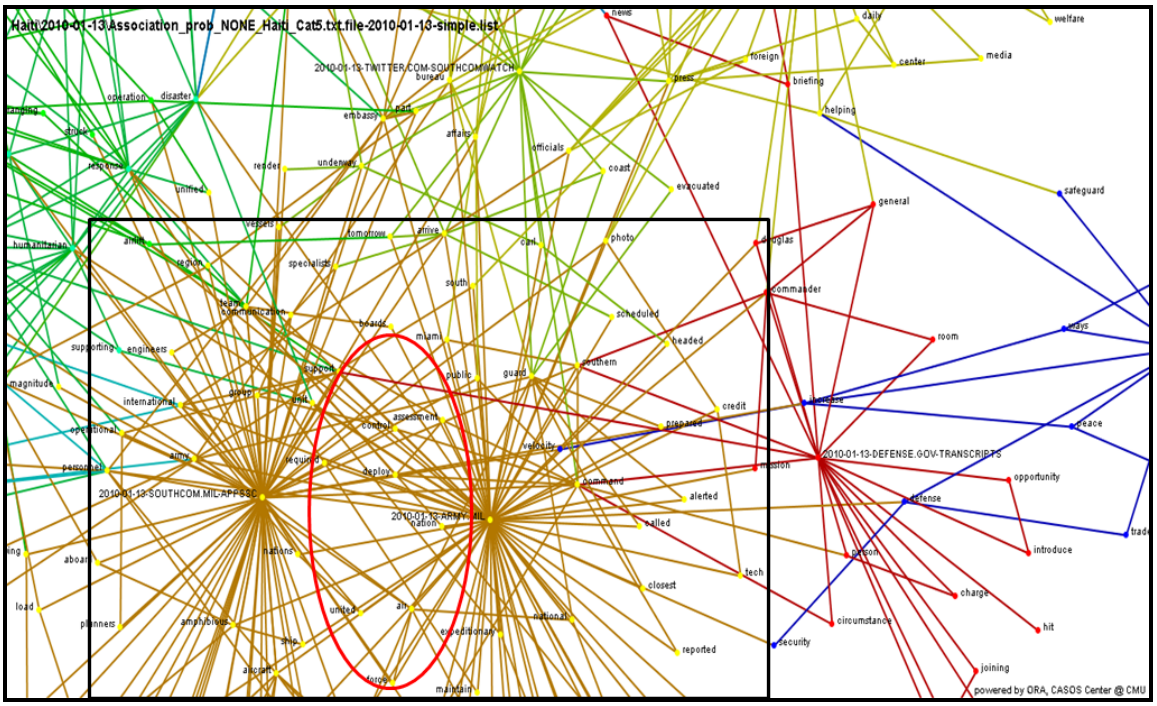

Figure 4(a): Synergy between SOUTHCOM and Army was high (same color)

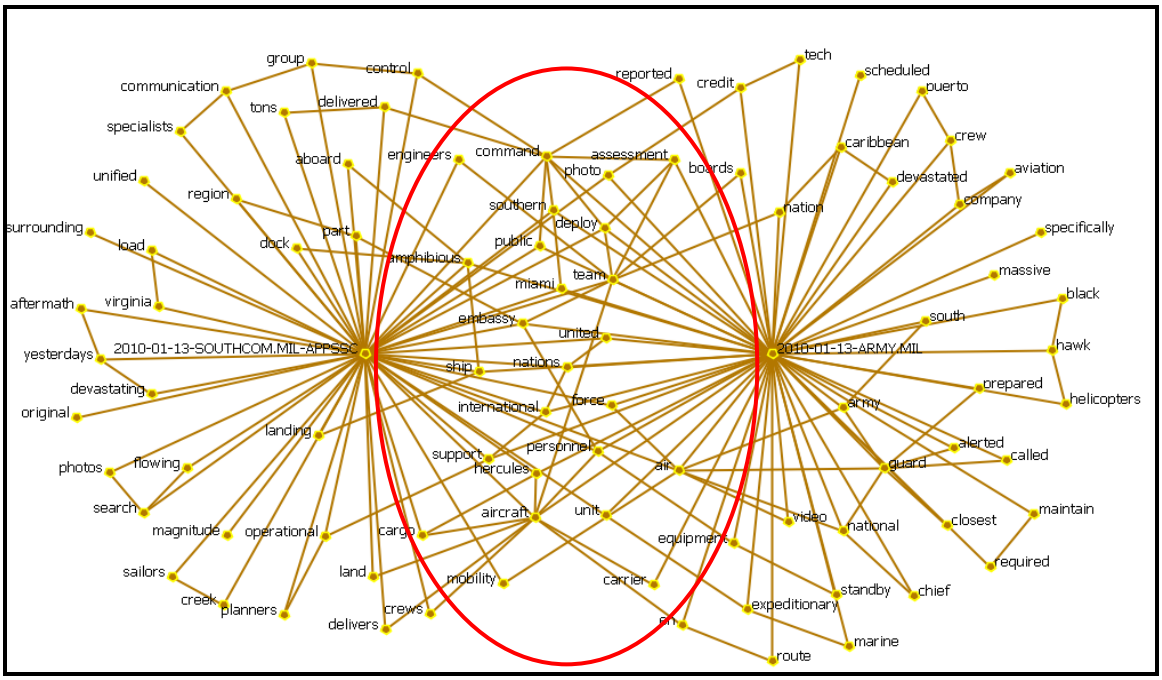

Figure 4(b): Details of the lexical links between SOUTHCOM and Army.

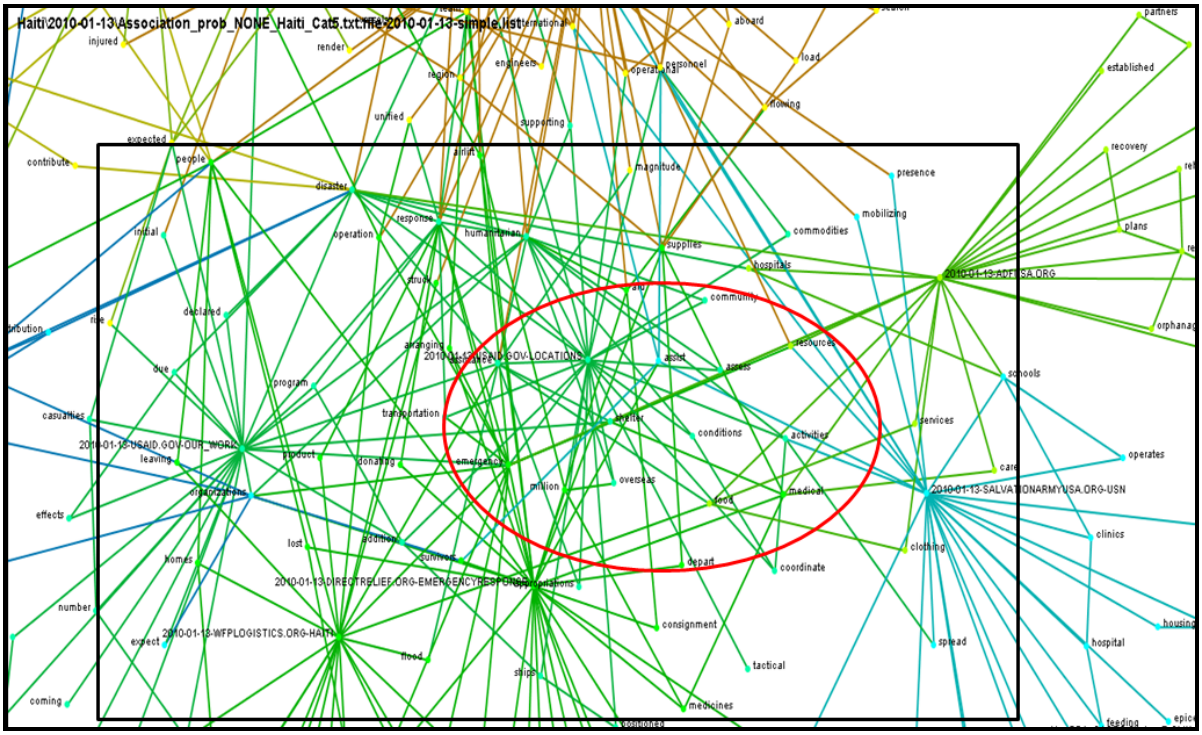

Figure 5(a): Synergy among various NGOs was also high.

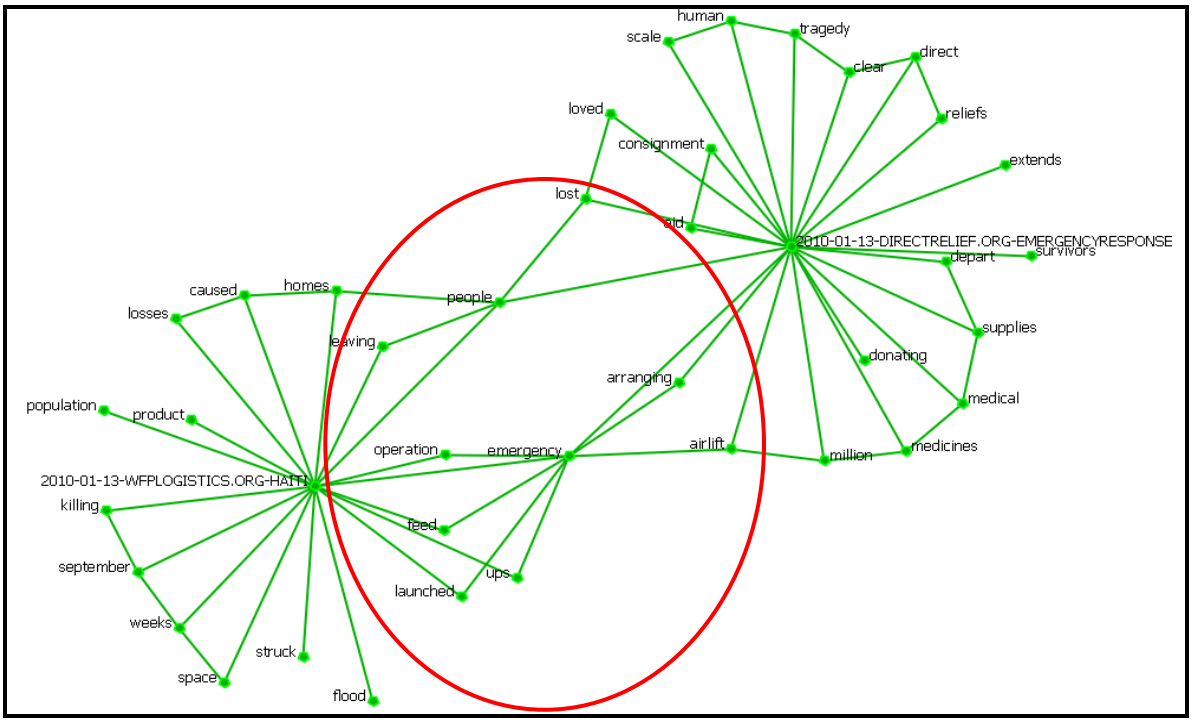

Figure 5(b): Lexical links between Direct Relief and WFP Logistics.

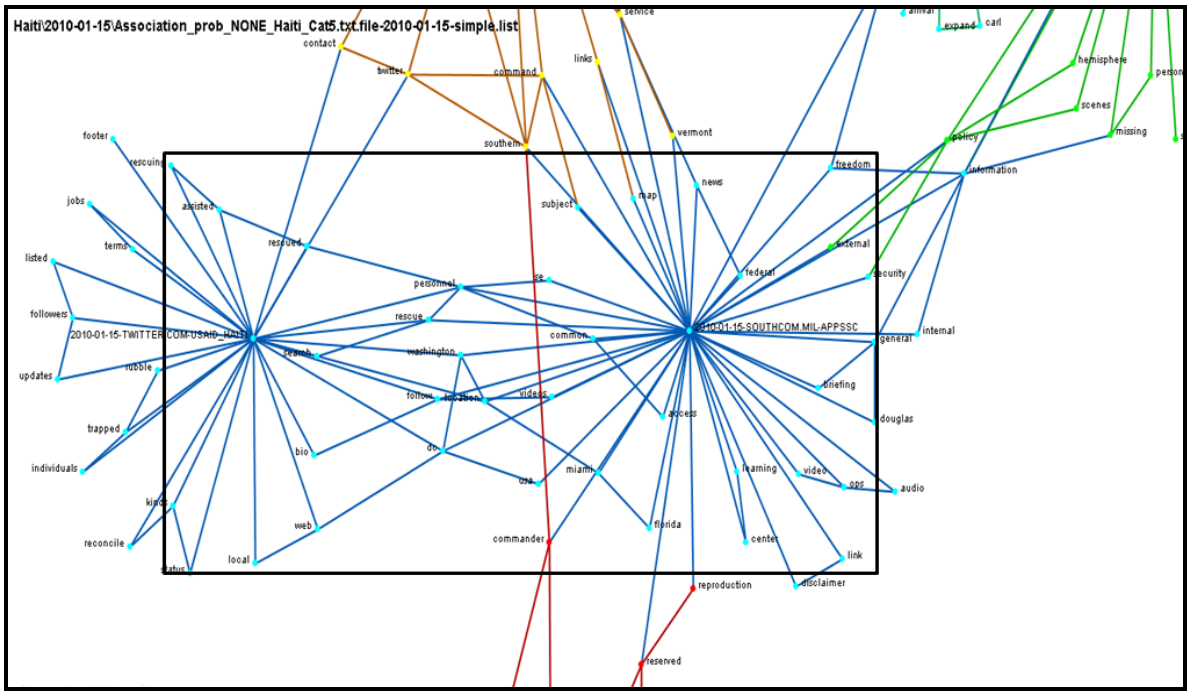

Figure 6(a): SOUTHCOM and USAID started to be synergetic on 2010-01-15.

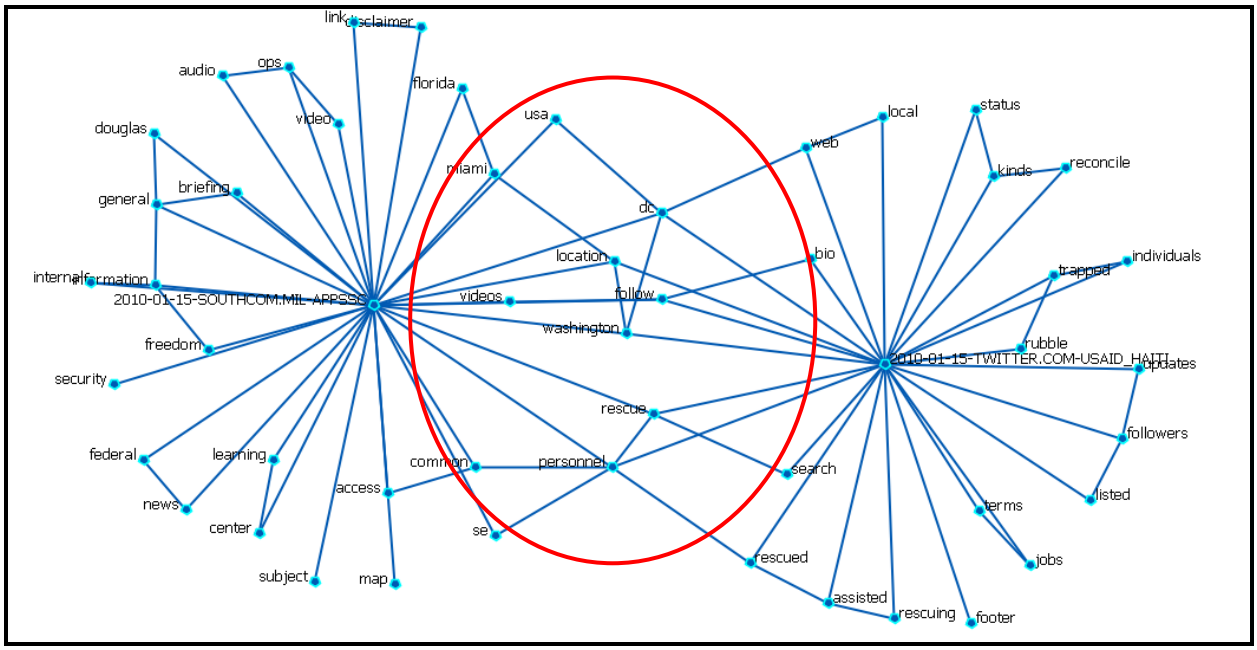

Figure 6(b): Lexical links between SOUTHCOM and USAID

Figure 6(a) shows, on 1/15/2010, SOUTHCOM and USAID started to develop synergy in the areas of “search”, “personnel rescue”, and “videos”, etc. Figure 6(b) shows an enlarged portion regarding the lexical links between SOUTHCOM and USAID.

4 Measures and Trends

As an overall view of the operation, Figure 7 shows the number of lexical links, number of nodes, number of word hubs among organizations and number of all words over time from 1/13 to 2/23. We found that:

- These measures were correlated.
- The number of lexical links (information) was large in the beginning, died down in the middle and went up again. This may correspond to different phases of the operation as highlighted in the keywords shown in the picture.

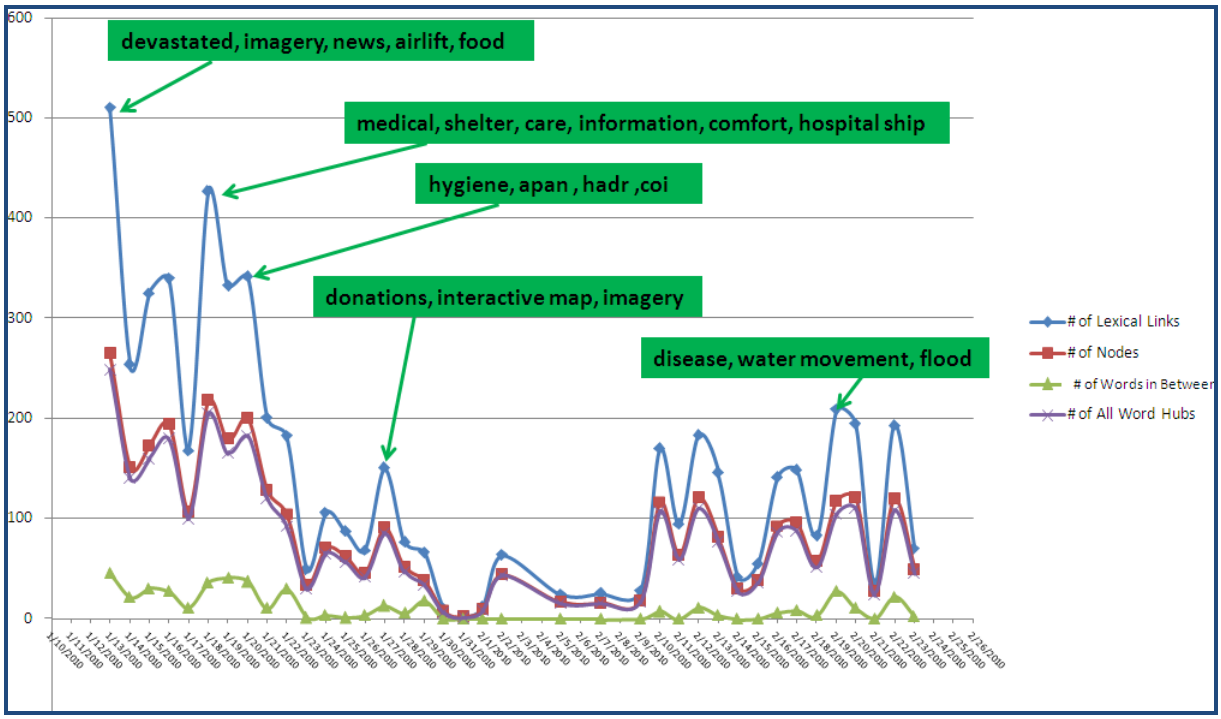

Figure 7: The number of lexical links (information) was large in the beginning, became smaller in the middle and went up again.

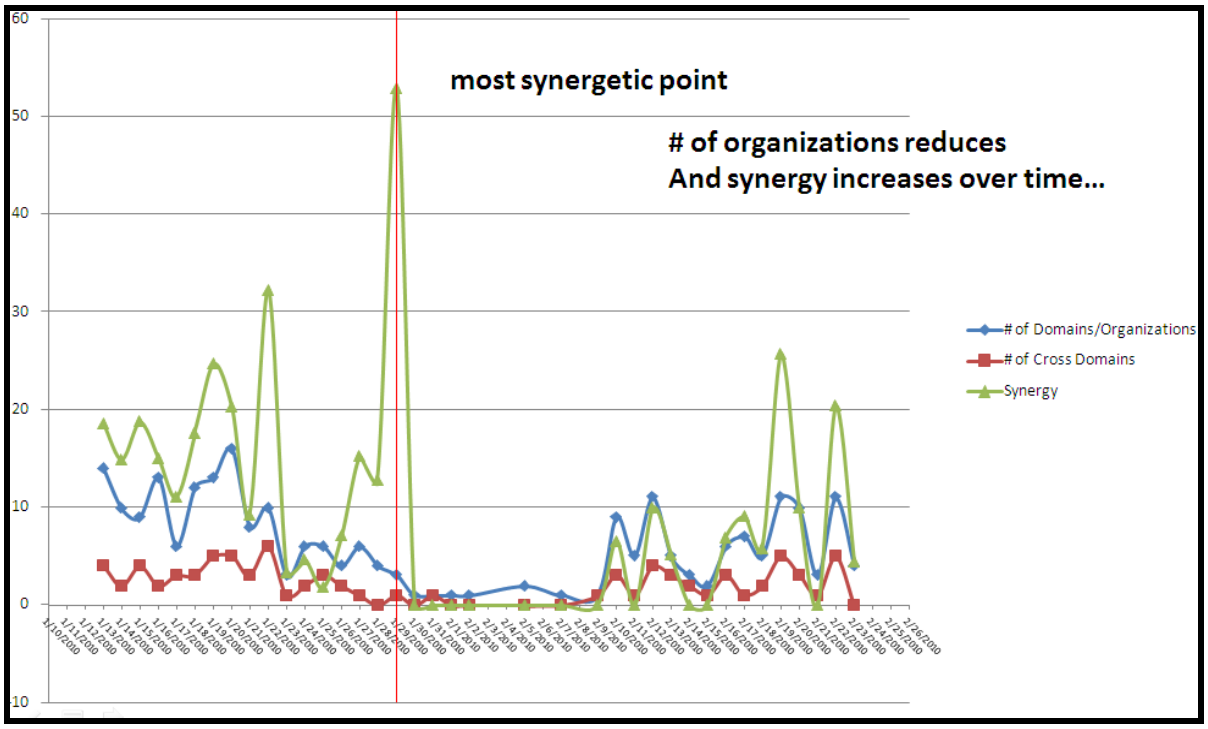

Figure 8: Trends of synergy and number of organizations

Figure 8 shows the number of organizations, number of cross-domain, and synergy index defined above over the same time period. We found that synergy was going up over time and inversely proportional to the number of organizations involved. On the day 1/29, the synergy was peaked and the number of organizations was minimized - it indicated the optimal efficiency or competency of the interagency operation reached at this point.

Appendix Table 1 lays out chronologically the organizations discovered in the open source data. Some observations from this table are listed as follows:

- The organizations were sorted by the time according to the first date when an organization showed up in the news. For example, southcom.mil and army.mil showed up on 1/13 and they remained there afterwards. "Inrelief.org" and "community.apan.org" only came up a week or so after the disaster on 1/19 and 1/20.
- Organizations were also divided into "military" and "other participating organizations" for the purpose of studying interagency operations.
- Cross-domain ones were included, for example twitter-southcomwatch or facebook.com-chiefofnavaloperations, showing SOUTHCOM and Chief of Naval Operations used Twitter and Facebook for the operation.

The sorted list provides insights for productive next steps toward our situational awareness of Haiti operations. This may, in turn, inform our theoretic understanding of how to guide future disaster efforts.

5 Conclusion

As we showed, using samples of open source information for the interagency collaboration of Haiti earthquake relief operations, that lexical links from LLA can be used to measure efficiency of interagency collaborations in a disaster relief operation. The number of overlapping lexical links can be used to measure synergy between two collaborating organizations. The synergy among the organizations in the Haiti relief operations was low in the beginning and became higher as time went on. We also found that social media Twitter, Facebook and Google provided critical capabilities for fast information gathering and dissemination for military entities such as SOUTHCOM. LLA provides a method to look at the interagency collaboration directly from the real-life communications and documentations.

6 Appendix

Table1: Organizations discovered in the open source data

	Military	**Other Participating Organizations**
1/13/2010		fieldnotes.unicefusa.org
	southcom.mil	
	army.mil	
		newsroom.redcross.org
	state.gov	
	twitter.com-southcomwatch	
		adfusa.org
		wfplogistics.org
		usaid.gov
		directrelief.org
		salvationarmyusa.org-usn
	defense.gov	
1/14/2010		
	ahp.us.army.mil	
		interaction.org
		haiti.usembassy.gov
		caritas.org
		actionaidusa.org
	dhs.gov	
1/15/2010		
		medicalteams.org-sf
	blogs.state.gov	

	haiticomfort.blogspot.com	
1/16/2010		
		msh.org
	marines.mil	
		sitroom.ushahididev.com
		tsa.gov
1/17/2010		
		twitter.com-usaid_haiti
	facebook.com-chiefofnavaloperations	
		airserv.org
	jtfb.southcom.mil	
1/18/2010		
	airforcelive.dodlive.mil	
	travel.state.gov	
	coastguardallhands.blogspot.com	
1/19/2010		
	twitter.com-usnscomfort	
		heartlandalliance.org
		arcrelief.org
	navy.mil-swf	
		inrelief.org
1/20/2010		
		cidi.org
		dec.usaid.gov
		kars.ku.edu
	facebook.com-usairforce	
		facebook.com-redcross
		community.apan.org
		wfp.org
	af.mil	
1/22/2010		
		hope140.org
	whitehouse.gov	
	slideshare.net-jtfhaiti	

1/24/2010		
	youtube.com-ussouthcom	
2/1/2010		
		oxfamamerica.org
2/6/2010		
		d7publicaffairs.com
2/9/2010		
	youtube.com-afbluetube	
2/10/2010		
	twitter.com-wfp	
		worldconcern.org
		dvidshub.net
		opusa.org
		un.org-apps
2/11/2010		
		mercycorps.org
		facebook.com-washingtonpostworld
2/12/2010		
		washingtonpost.com-wp
		helpageusa.org
		stophungernow.org
	facebook.com-jtfhaiti	
	twitter.com-ushahidi	
	hhs.gov	
		afsc.org
		globalfoodforthought.typepad.com
2/13/2010		
	facebook.com-ussvinson	
	uscg.mil-comdt	
2/14/2010		
		yele.org
2/15/2010		
		twitter.com-haiti_inrelief
2/16/2010		

	facebook.com-uscg.d7.publicaffairs	
	fpc.state.gov	
		www2.nbc13.com-vtm
	coastguard.dodlive.mil	
2/17/2010		
		crs.org
		worldconcern.org
		chfinternational.org
2/18/2010		
		washingtonpost.com-wp
	twitter.com-jtfhaiti	
2/19/2010		
		csmonitor.com
		twitpic.com
		new.gbgm-umc.org
2/20/2010		
	army.mil-defensemediaactivity	
		reliefweb.int-rw
		ochaonline.un.org
		un.org
		oxfam.org
		ireport.com
2/23/2010		
		internationalchildcare.org
		stripes.com
		globallinks.org

7 References

AutoMap. (2009). AutoMap**:** Extract, Analyze and Represent Relational Data from Texts. Retrieved from http://www.casos.cs.cmu.edu/projects/automap/

Blei, D., Ng, A. & Jordan, M. (2003). Latent Dirichlet Allocation. Journal of Machine Learning Research 3, 993–1022. Retrieved from http://jmlr.csail.mit.edu/papers/volume3/blei03a/blei03a.pdf.

Foltz, P. W. (2002). Quantitative cognitive models of text and discourse processing. In *The handbook of discourse processes*. Mahwah, NJ: Lawrence Erlbaum Publishing.

Gallup, S. P., MacKinnon, D. J., Zhao, Y., Robey, J., & Odell, C. (2009, October 6–8). Facilitating decision making, re-use and collaboration: A knowledge management approach for system self-awareness. Paper presented at the International Joint Conference on Knowledge Discovery, Knowledge Engineering, and Knowledge Management (IC3K), Madeira, Portugal.

Girvan, M., & Newman, M. E. J. (2002, June). Community structure in social and biological networks. In the Proceedings of the National Academy of Sciences of the United States of America, Vol. 99, No. 12,, p. 7821-7826.

Gerber, C. (2005). Smart searching, new technology is helping defense intelligence analysts sort through huge volumes of data. *Military Information Technology*, *9*(9). Retrived from http://www.mkbergman.com/171/large-scale-intelligence-analysis/

Lexical Analysis (LA), (2010). Lexical Analysis. Retrieved from http://en.wikipedia.org/wiki/Lexical_analysis

Lewis, K., Kaufmana, J., Gonzaleza, M., Wimmerb, A. & Christakis, N.(2008). "Tastes, Ties, and Time (T3): A new social network dataset using Facebook.com", *Social Networks* 30 (2008) 330–342

Quantum Intelligence,(QI). (2009). Quantum Intelligence, Inc. Retrieved from http://www.quantumii.com

Schensul, J. J., Schensul, S. L., & LeCompte, M. D. (1999). Essential ethnographic methods: Observations, interviews and questionnaires. Lanham, MD: Rowman Altamira.

Tecuci, G., Boicu, M., Marcu, D., Boicu, C., Barbulescu, M., Ayers, C., & Cammons, D. (2007). Cognitive assistants for analysts. In J. Auger & W. Wimbish (Eds.), Carlisle Barracks:National Intelligence University, Office of the Director of National Intelligence, and U.S. Army War College Center for Strategic Leadership.

Wasserman, S., &Faust, K. (1994). *Social network analysis: Methods and applications*. Cambridge, England: Cambridge University Press.

WordNet (2009) http://wordnet.princeton.edu/

Zhao, Y., Gallup, S., & MacKinnon, D. (2010). Towards real-time program awareness via lexical link analysis. In *Proceedings of the Seventh Annual Acquisition Research Program*. Monterey, CA: Naval Postgraduate School.