SYM-AM-15-108

PROCEEDINGS
OF THE
TWELFTH ANNUAL ACQUISITION RESEARCH SYMPOSIUM

THURSDAY SESSIONS
VOLUME II

Lexical Link Analysis Application: Improving Web Service to Acquisition Visibility Portal Phase III

Ying Zhao, NPS
Doug MacKinnon, NPS
Shelley Gallup, NPS

Published April 30, 2015

ACQUISITION RESEARCH PROGRAM
GRADUATE SCHOOL OF BUSINESS & PUBLIC POLICY
NAVAL POSTGRADUATE SCHOOL

The research presented in this report was supported by the Acquisition Research Program of the Graduate School of Business & Public Policy at the Naval Postgraduate School.

To request defense acquisition research, to become a research sponsor, or to print additional copies of reports, please contact any of the staff listed on the Acquisition Research Program website (www.acquisitionresearch.net).

Lexical Link Analysis Application: Improving Web Service to Acquisition Visibility Portal Phase III

Ying Zhao—is a research associate professor at the Naval Postgraduate School and frequent contributor to DoD forums on knowledge management and data sciences. Her research and numerous professional papers are focused on knowledge management approaches such as data/text mining, Lexical Link Analysis, system self-awareness, Collaborative Learning Agents, search and visualization for decision-making, and collaboration. Dr. Zhao was principal investigator (PI) for six contracts awarded by the DoD Small Business Innovation Research (SBIR) Program. Dr. Zhao is a co-author of four U.S. patents in knowledge pattern search from networked agents, data fusion, and visualization for multiple anomaly detection systems. She received her PhD in mathematics from MIT and is the co-founder of Quantum Intelligence, Inc. [yzhao@nps.edu]

Doug MacKinnon—is a research associate professor at the Naval Postgraduate School. MacKinnon is the deputy director of the Distributed Information and Systems Experimentation (DISE) research group where he leads multidisciplinary studies ranging from leading the Analyst Capability Working Group for the U.S. Air Force and studying Maritime Domain Awareness, as well as Knowledge Management (KM) and Lexical Link Analysis projects. He also led the assessment for the Tasking, Planning, Exploitation, and Dissemination process during the Empire Challenge 2008 and 2009 (EC08/09) field experiments and for numerous other field experiments of new technologies during Trident Warrior 2012. Dr. MacKinnon teaches courses in operations research and holds a PhD from Stanford University, conducting successful theoretic and field research in KM. He has served as the program manager for two major government projects of over $50 million each, implementing new technologies while reducing manpower requirements. He has served over 20 years as a naval surface warfare officer, amassing over eight years at sea and serving in four U.S. Navy warships with five major underway deployments. [djmackin@nps.edu]

Shelley Gallup—is a research associate professor at the Naval Postgraduate School's Department of Information Sciences and the director of Distributed Information and Systems Experimentation (DISE). Dr. Gallup has a multidisciplinary science, engineering, and analysis background, including microbiology, biochemistry, space systems, international relations, strategy and policy, and systems analysis. He returned to academia after retiring from naval service in 1994 and received his PhD in engineering management from Old Dominion University in 1998. Dr. Gallup joined NPS in 1999, bringing his background in systems analysis, naval operations, military systems, and experimental methods first to the Fleet Battle Experiment series (1999–2002) and then to Fleet experimentation in the Trident Warrior series (2003–2013). Dr. Gallup's interests are in knowledge management and complex systems field experimentation. [spgallup@nps.edu]

Abstract

We have been studying DoD acquisition decision-making since 2009. The U.S. DoD acquisition process is extremely complex. There are three key processes that must work in concert to deliver capabilities: determining warfighters' requirements/needs, the DoD budget planning, and the procurement of final products. Each process produces large amounts of information (Big Data). There is a critical need for automation, validation, and discovery to help acquisition professionals, decision-makers, and researchers understand the important content within large data sets and optimize DoD resources throughout the processes. Lexical Link Analysis (LLA) can help, by applying automation to reveal and depict—to decision-makers—the correlations, associations, and program gaps across all or subsets of acquisition programs over many years. This enables strategic understanding of data gaps and potential trends, and can inform managers where areas might have higher program risk and how resource and big data management might affect the desired return on investment among projects. In this paper, we describe new developments in analytics and visualization, how LLA is adaptive to Big Data Architecture and Analytics (BDAA), and needs for Big Acquisition Data used in Defense Acquisition Visibility Environment (DAVE).

Background

We have been studying Department of Defense (DoD) acquisition decision-making since 2009 (Gallup et al., 2009; Zhao, Gallup, & MacKinnon, 2010, 2011, 2012a, 2012b, 2013, 2014). The U.S. DoD acquisition process is extremely complex. There are three key processes that must work in concert to deliver capabilities: definition of warfighters' requirements/needs, DoD budget planning, and procurement of products, as in Figure 1. Each process produces volumes of information (Big Data). The need for automation, validation, and discovery is now a critical need, as acquisition professionals, decision-makers, and researchers grapple to understand data and make decisions to optimize DoD resources.

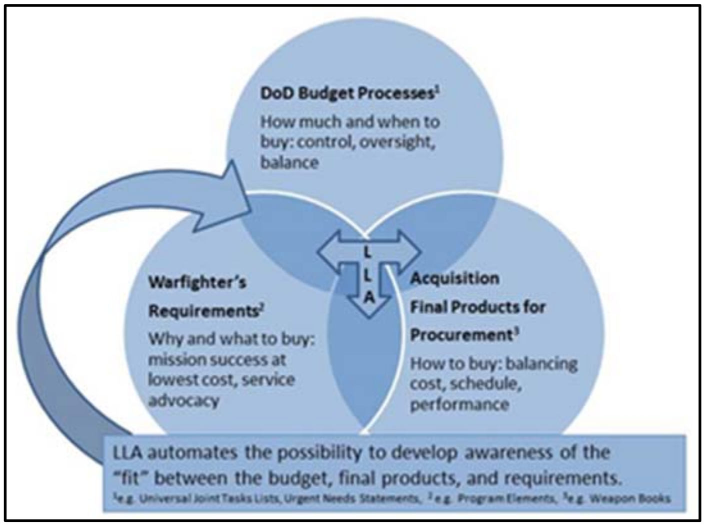


Figure 1. DoD Acquisition Decision-Making

Since 2009, we have been working on the problem of how the interlocking systems processes become aware of their fit between DoD programs and warfighters' needs. How are gaps revealed? Moreover, in the performance of DoD acquisition processes, each functional community is required to review only the particular information for which it is responsible, further exacerbating the problem of lack of fitness. For example, the systems engineering community typically only examines the engineering documents and feasibility studies, the test and evaluation community looks only at the test and evaluation plans, and the acquisition community looks at the acquisition strategies. Rarely do these stakeholders review each other data or jointly discuss the core questions and integrated processes together as shown in Figure 1.

Motivated by this lack of fit and horizontal integration, we have been applying Lexical Link Analysis (LLA), a data-driven automation technology and methodology across DoD acquisition processes to

- surface themes and their relationships across multiple data sources
- discover high value areas for investment
- compare/correlate data from multiple data sources
- sort/rank important and interesting information

LLA is a data-driven method for pattern recognition, anomaly detection, and data fusion. It shares indexes not data, feasible for parallel and distributed processing, adaptive to Big Data Architecture and Analytics (BDAA) and needs for Big Acquisition Data.

As an example from past work, we took a detailed look at the Research, Development, Test and Evaluation (RDT&E) budget modification practice from one year to the next over the course of 10 years and about 450 DoD program elements. We found a pattern that the programs with fewer links (measured by LLA) to warfighters' requirements received more budget reduction in total but less on average, indicating the budget reduction may have focused only on large and expensive programs rather than perhaps cutting all the programs that do not match warfighters' requirements. Furthermore, the programs with more links to each other received more budget reduction in total, as well as on average, indicating a pattern of good practice of allocating DoD acquisition resources to avoid overlapping efforts and to fund new and unique projects. These findings were useful as validation and guidance for future decision processes for automatically identifying programs to match the warfighter's requirements, limit overall spending, minimize efficiencies, eliminate unnecessary cost, and maximize the return on investment.

In this paper, we demonstrate a set of comprehensive LLA analysis reports and visualizations generated automatically from multiple data sources. These reports and visualizations reveal data correlations and gaps among multiple data sources. These correlations and gaps could form the basis for pattern recognition, anomaly detection, and further inquiry or future reconciliation of the expectations (e.g., acquisition strategy) and realities (e.g., engineering feasibility) from various communities. The automatic discovery of the disconnection or gaps could be fed back to the human analysts or decision-makers for decision-making and resource management.

Methodology

Lexical Link Analysis (LLA)

LLA has been used to analyze unstructured and structured data for pattern recognition, anomaly detection, and data fusion. It uses the theory of system self-awareness (SSA) to identify high-value information in the data that can be used to guide future decision processes in a data-driven or unsupervised learning fashion. It is implemented via a smart infrastructure named "system and method for knowledge pattern search from networked agents (U.S. patent 8,903,756)," also known as Collaborative Learning Agents (CLA), licensed from Quantum Intelligence, Inc. (Zhou, Zhao, & Kotak, 2009).

In LLA, a complex system is expressed in specific vocabularies or lexicons to characterize its features, attributes, or surrounding environment. LLA uses bi-gram word pairs as the features to form word networks. Figure 2 depicts using LLA to analyze 10 years of reports in the Naval Postgraduate School (NPS) Acquisition Research Program with word pairs as groups or *themes*. Figure 3 shows a detail of a *theme* in Figure 2. A node represents a word. A link or edge represents a word pair.

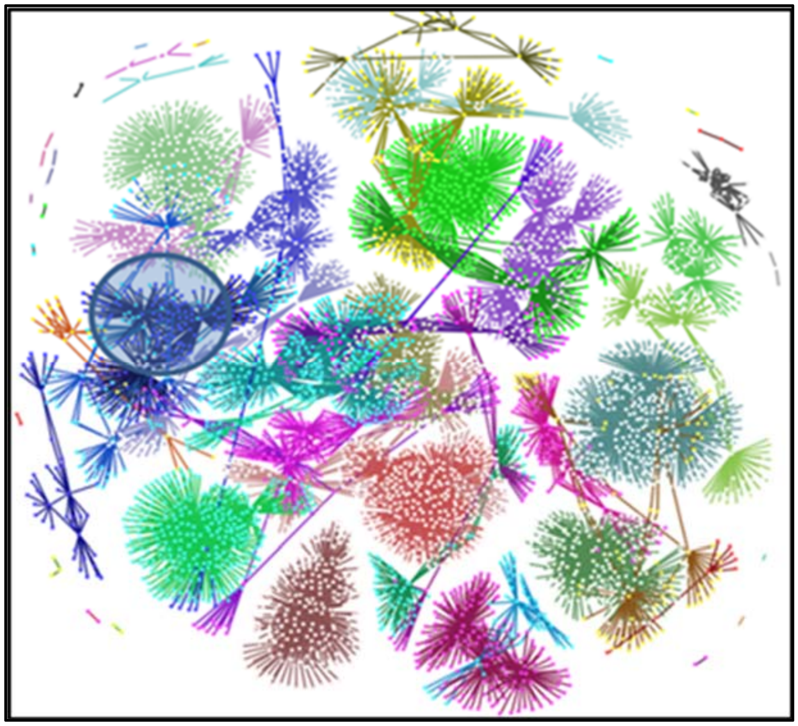


Figure 2. Themes Discovered in Colored Groups

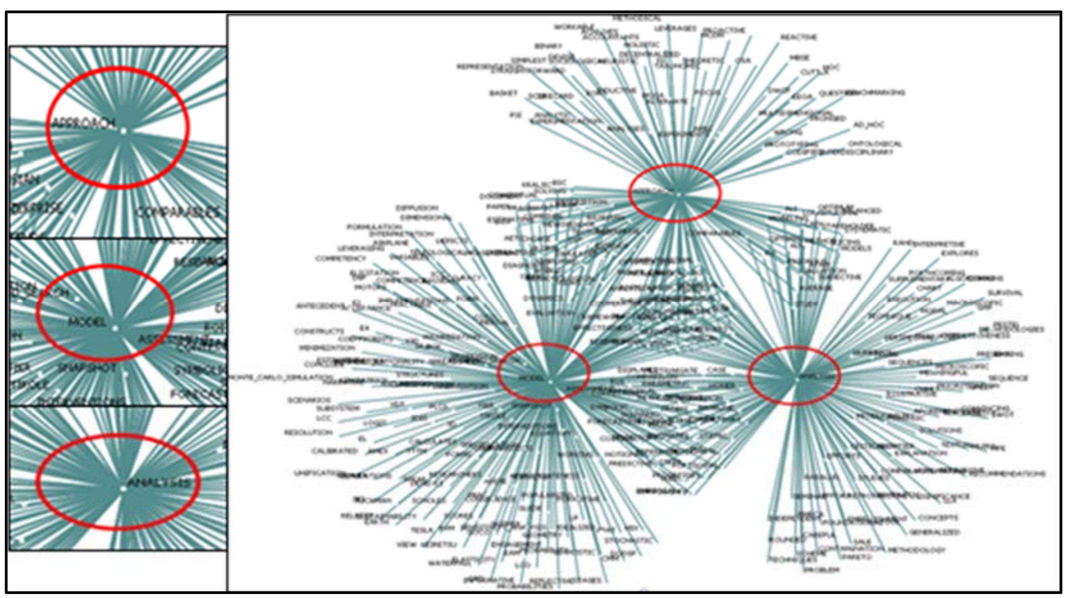


Figure 3. A Detailed View of a Theme in Figure 2

LLA is related to bags-of-words (BAG) methods such as LDA (Blei, Ng, & Jordan, 2003) and text-as-network (TAN) methods such as the Stanford Lexical Parser (SLP; Stanford Natural Language Processing Group [SNLPG], 2015). LLA selects and groups features into three basic types:

- Popular (P): They are the main themes in the data. Figure 3 is an example of a popular theme centered around word nodes “analysis,” “model,” and “approach.” These themes could be less interesting because they are already in the public consensus and awareness. They represent the patterns in the data.

- Emerging (E): Themes may grow to be popular over time. Figure 4 is an example of an emerging theme centered around word nodes “national,” “defense,” and “acquisition.”
- Anomalous (A): These themes may be off-topic themes that are interesting for further investigation. Figure 5 is an example of anomalous theme centered around word nodes “stock” and “market(s).”

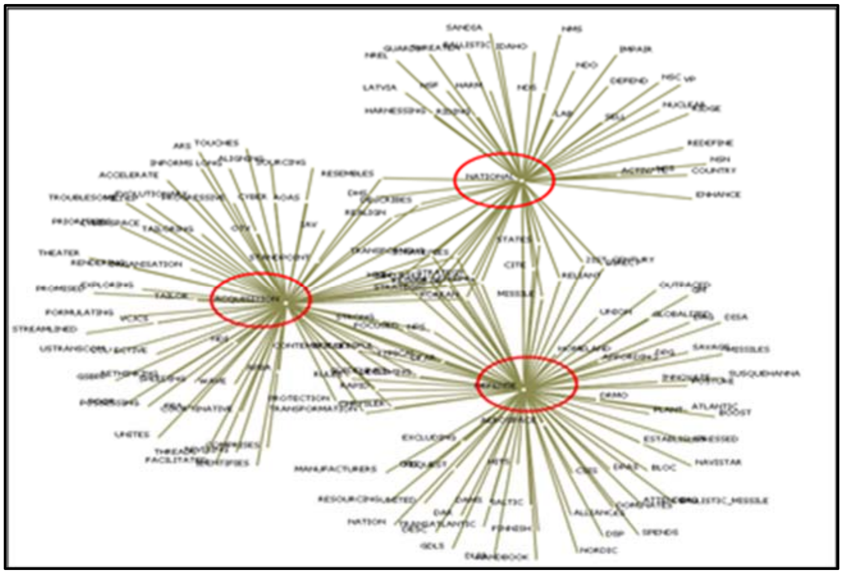


Figure 4. An Example of Emerging Theme

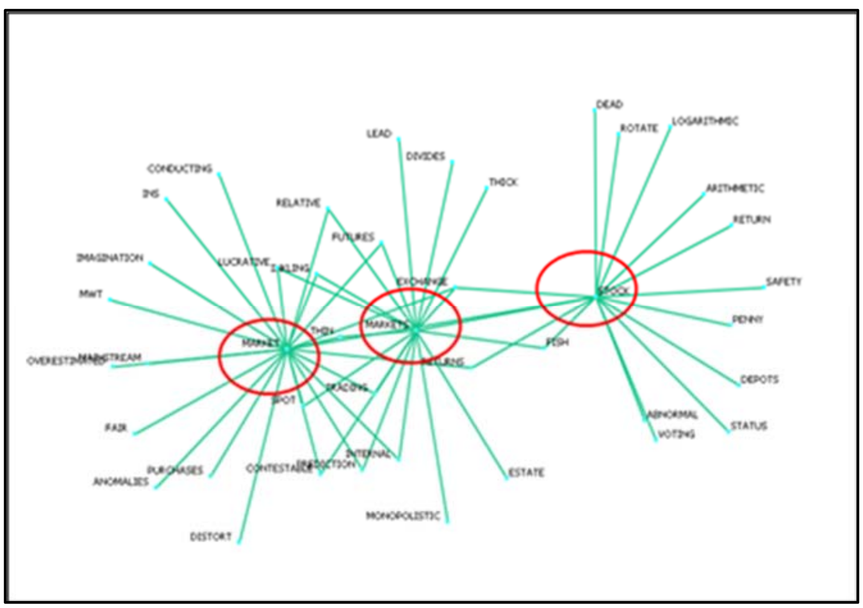


Figure 5. An Example of Anomalous Theme

Figure 6 summarizes LLA used for historical and new data. The red part shows a pattern (e.g., a theme) discovery phase using historical data including data fusion that come from multiple learning agents. The black part shows an application phase that new data is compared with the patterns discovered and hence the anomalies are revealed.

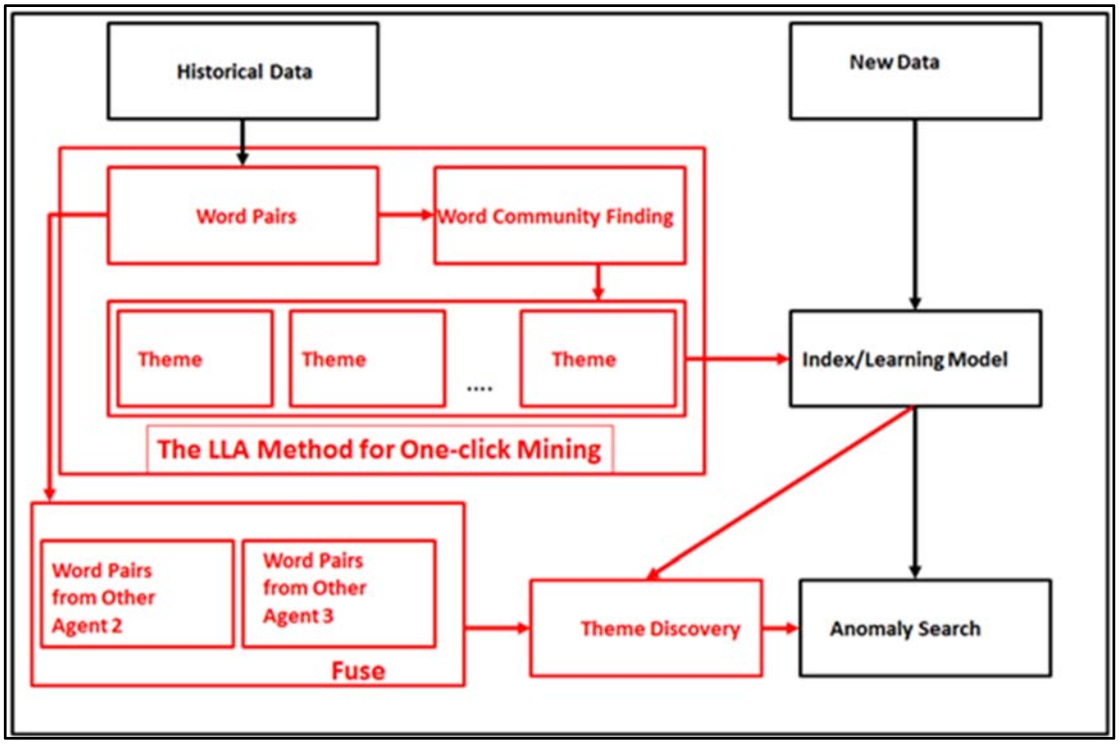


Figure 6. Diagram for the LLA Method

Word Pairs Generalization and CCC Method

Figure 7 shows the word pairs/bi-gram in an LLA can be generalized as a Context-Concept-Cluster (CCC) model, where a context is a generic attribute that can be shared by multiple data sources, a concept is a specific attribute for a data source, and a cluster is a combination of attributes or themes that can be computed using a word community finding algorithm (e.g., Girvan & Newman, 2002) in Figure 6 to characterize a data set. Context can be a word, location, time, or object, and so on.

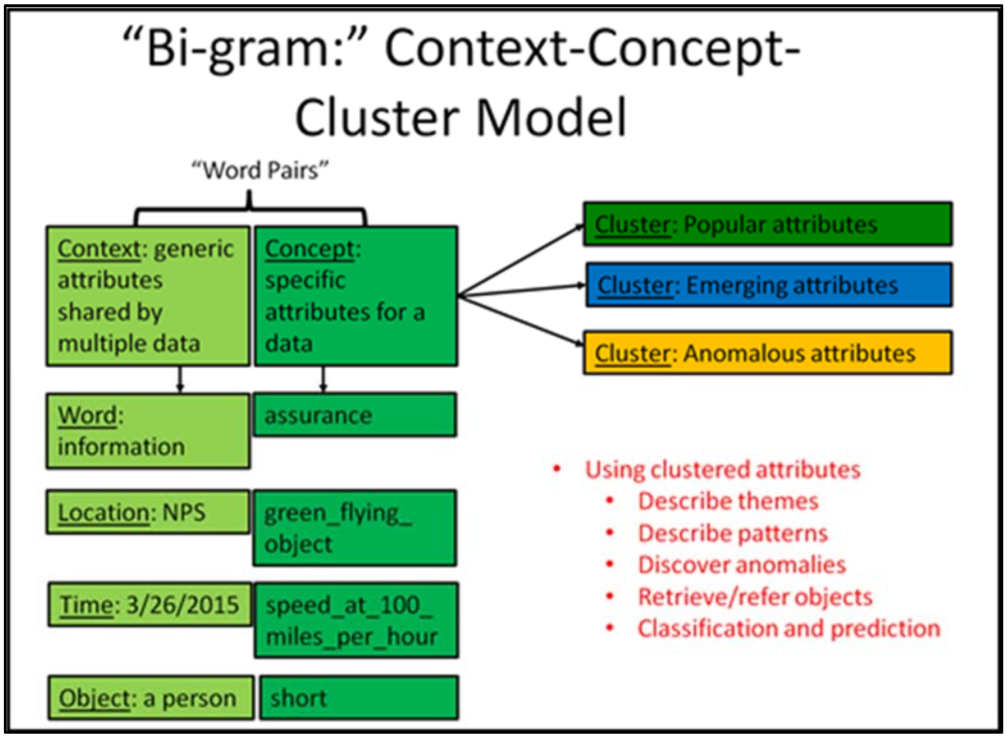


Figure 7. The Word Pairs/Bi-Gram in LLA a Context-Concept-Cluster (CCC) Model

Figure 8 summarizes how a generalized CCC method is used for historical and new data. Similar to Figure 6, there is a pattern discovery phase using historical data where patterns are learned and discovered, and an application phase for a new data is compared with the patterns discovered, and anomalies are revealed.

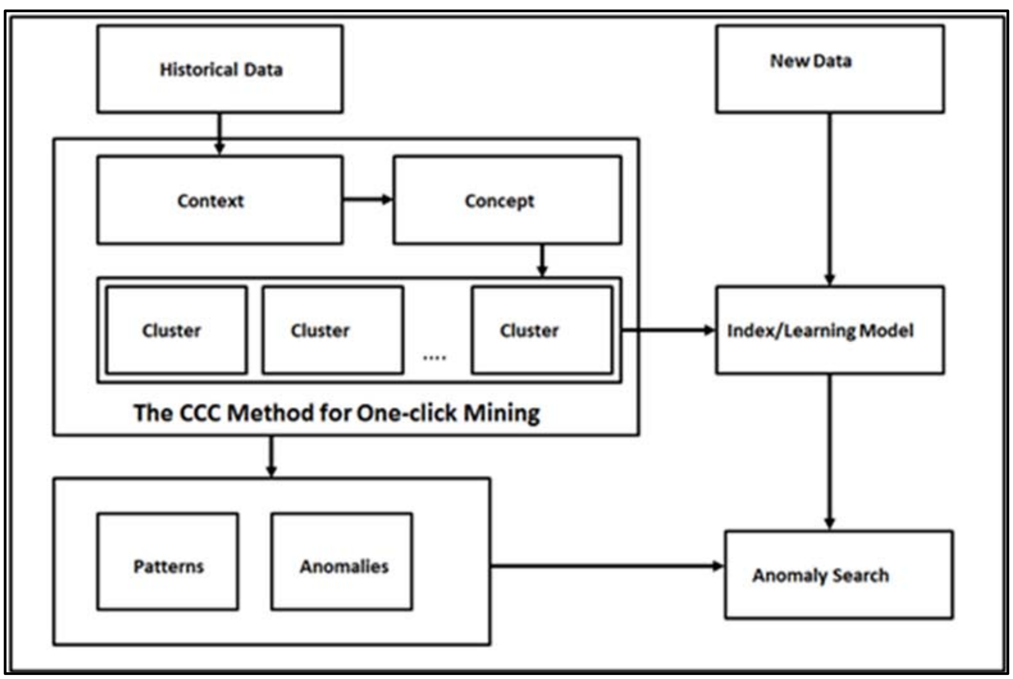


Figure 8. Diagram for the CCC Model

Research Results

Task 1

We are working with the OSD OUSD(ATL) (US) to install the LLA/SSA/CLA system as a web service using a Linux platform (i.e., CentOS) in the Defense Acquisition Visibility Environment (DAVE) test bed. We created a publically available data set with the installation to test. In this example, data sources include 10 days of business news of about 1,000 companies, which are organized in industries as follows:

- Technology
- Services
- Healthcare
- Utilities
- Basic Materials
- Financial
- Consumer Goods
- Industrial Goods
- Conglomerates

Each category of information such as “Healthcare” or “Consumer Goods” are indexed, mined, and listed under “Index Management” separately in a single LLA server. When clicking “Fuse,” these indexed/mined models are fused into one model. Figure 9 shows “Fuse Results” from LLA listed.

Figure 9. Fuse Results Listed

Figure 10 shows the discovered themes, where green themes 101(P) and 20(P) are “popular” themes, blue themes 156(E), 49(E), and 46(E) are “emerging” themes, and gold themes 208(A), 62(A), and others are “anomalous” themes.

- Popular themes are the main themes in the data. Figure 11 is an example of a popular theme centered “dividend cuts, see dividend” for this data. Columns “Consensus” is the ratio of the number of matched word pairs (i.e., at least two data sources contain the word pairs) over the number of unique word pairs (i.e., only one data source contains the word pairs). These themes could be less interesting because they are already in the public consensus and awareness. They represent the patterns in the data. The red links represent the word pairs that are shared for at least two data sources while the black data sources are unique to one data source.

- Emerging themes may grow to be popular over time. Figure 12 is an example of an emerging theme centered “back shares, Canada back.”
- Anomalous themes may be off-topic themes that are interesting for further investigation. Figure 13 is an example of anomalous theme centered around “top buys, set top.” Anomalous concepts are more interesting to investigate, for example, concepts in Figure 13 such as “buys web,” “streaming service,” “buys insider,” “web ipo,” and so on, may have better returns on investment than the concepts in a popular theme such as “sees dividend” and “announces positive.”

Discovered Themes

Event_Date_Sort	Theme Id	Theme Keywords	Visualization	Matched	Unique	Total	Consensus	Gaps
all	101(P)	DIVIDEND CUTS,SEES DIVIDEND	Vis	494	67	561	0.88	0.12
all	20(P)	LLC ANNOUNCES,ANNOUNCES POSITIVE	Vis	213	74	287	0.74	0.26
all	156(E)	BACK SHARES,CANADA BACK	Vis	143	24	167	0.86	0.14
all	49(E)	HELPS COMPANIES,START HELPS	Vis	111	44	155	0.72	0.28
all	46(E)	COMPANY NAMED,EXECUTIVE NAMED	Vis	114	22	136	0.84	0.16
all	208(A)	TOP BUYS,SET TOP	Vis	128	32	160	0.80	0.20
all	62(A)	EXPANDS HEALTH,DATA EXPANDS	Vis	60	14	74	0.81	0.19
all	51(A)	PLANS CUSTOMER,IPO PLANS	Vis	77	17	94	0.82	0.18
all	182(A)	SECURITIES CALIFORNIA,MORTGAGE SECURITIES	Vis	76	18	94	0.81	0.19
all	44(A)	CAPITAL FUND,FUND RATING	Vis	103	20	123	0.84	0.16
all	100(A)	CITY CENTER,EUROPEAN CENTER	Vis	57	17	74	0.77	0.23
all	221(A)	UPDATE SUNOCO,UPDATE HR	Vis	46	9	55	0.84	0.16
all	53(A)	HIGHLIGHTS CONOCOPHILLIPS,HIGHLIGHTS HSBC	Vis	68	12	80	0.85	0.15
all	109(A)	FOOD INDUSTRY,INDUSTRY LEADING	Vis	38	12	50	0.76	0.24
all	191(A)	LEAD COUNTRY,SYSTEM LEAD	Vis	31	7	38	0.82	0.18
all	132(A)	WEEKLY MARKET,WEEKLY RECAP,MARKET RECAP	Vis	61	8	69	0.88	0.12
all	180(A)	CARD DEALS,GIFT CARD,GIFT CARD	Vis	39	12	51	0.76	0.24

Figure 10. Discovered Themes Listed

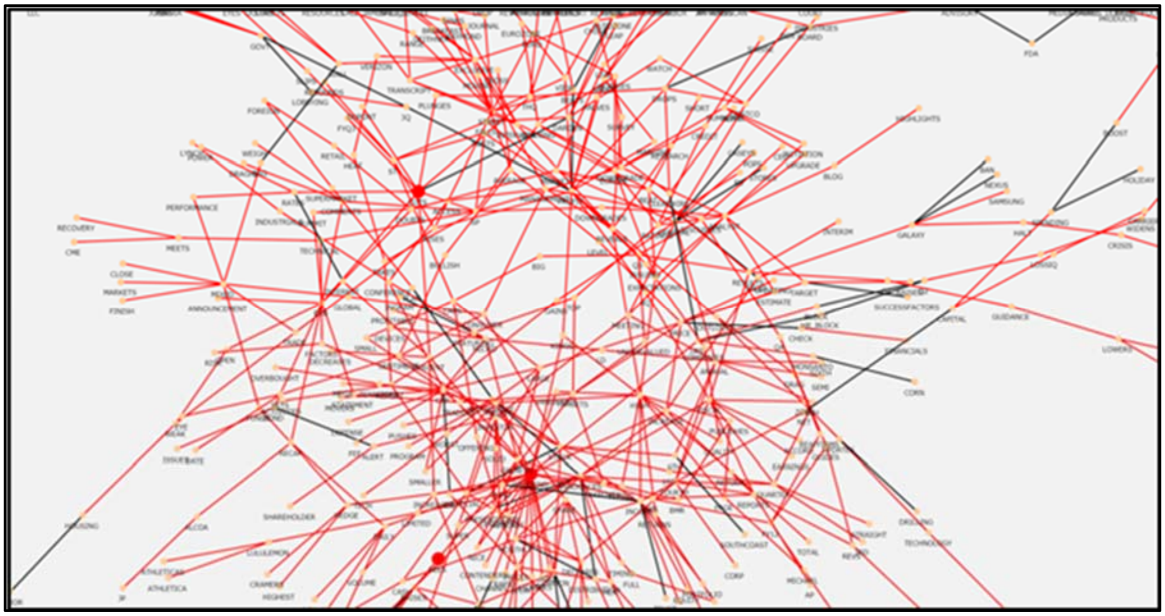


Figure 11. Visualization for the Popular Theme 101(P)

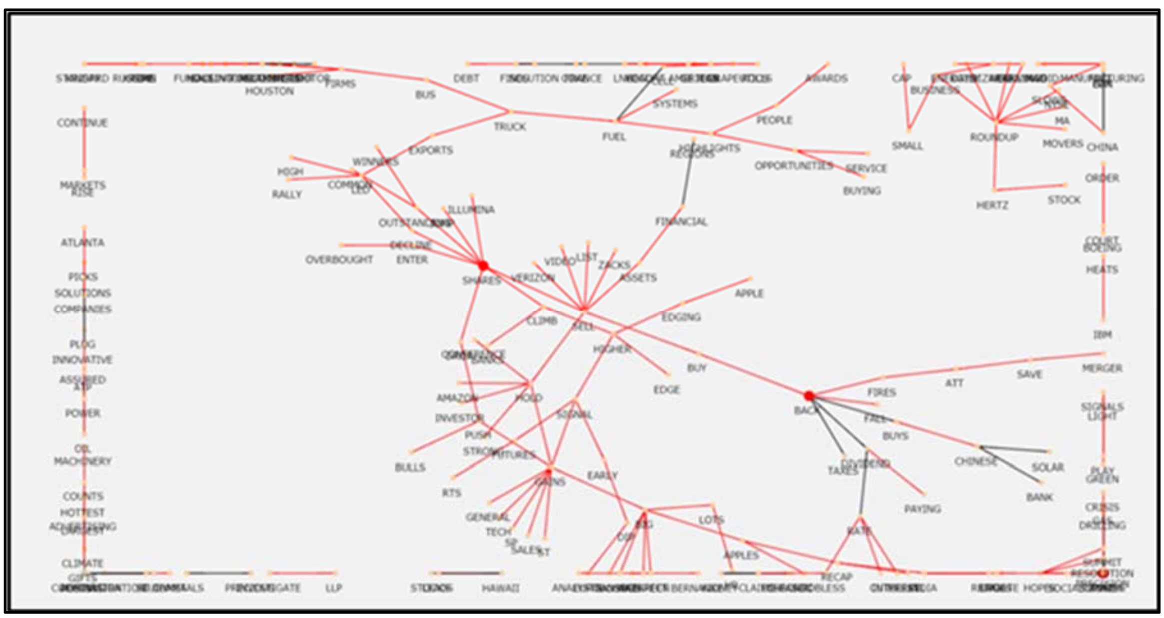


Figure 12. Emerging Themes (e.g., 156[E])

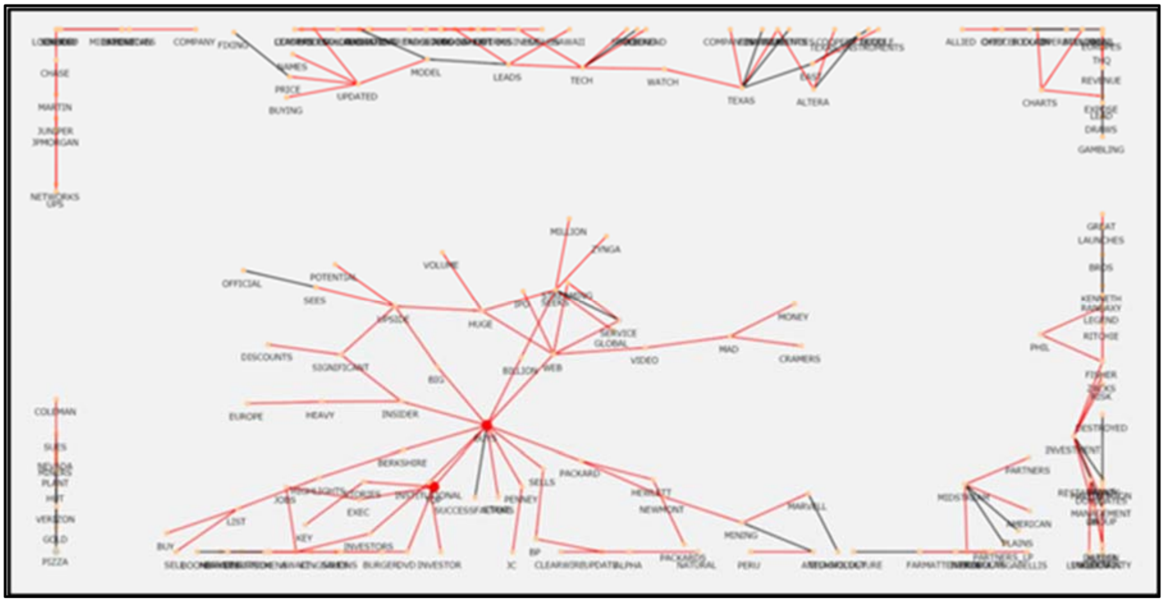


Figure 13. Anomalous Theme (e.g., 208[A])

Match Matrix Unique Word Pairs by Theme

Figure 14 shows the numbers of unique word pairs in a data source and a theme. For example, there are 12 unique word pairs for the data source "Index_BasicMaterials" in the theme titled "101:dividend cuts, sees dividend" in Figure 14. Clicking this number leads to a list showing the 12 word pairs as shown in Figure 15. Figure 16 shows the list can be further drilled down to a search result list (e.g., "sees energy") and to the original documents that contain the word pair.

Match Matrix From Lexical Link Analysis: Updated on Mon Mar 16 15:50:48 2015 Using 'Combined' Unique Word Pairs by Source

		Match Score	Uniqueness Score	101:DIVIDEND CUTS,SEES DIVIDEND	20:LLC ANNOUNCES,ANNOUNCES POSITIVE	101:DIVIDEND CUTS,SEES DIVIDEND	20:LLC ANNOUNCES,ANNOUNCES POSITIVE	101:DIVIDEND CUTS,SEES DIVIDEND	20:LLC ANNOUNCES,ANNOUNCES POSITIVE	101:DIVIDEND CUTS,SEES DIVIDEND	20:LLC ANNOUNCES,ANNOUNCES POSITIVE
1	Index_BasicMaterials	30.00	223.00	12.00	31.00	12.00	31.00	12.00	31.00	12.00	31.00
2	Index_Financial	28.00	270.00	21.00	29.00	21.00	29.00	21.00	29.00	21.00	29.00
3	Index_Services	27.00	328.00	30.00	26.00	30.00	26.00	30.00	26.00	30.00	26.00
4	Index_ConsumerGoods	18.00	120.00	19.00	6.00	19.00	6.00	19.00	6.00	19.00	6.00
5	Index_Technology	14.00	229.00	12.00	16.00	12.00	16.00	12.00	16.00	12.00	16.00
6	Index_IndustrialGoods	12.00	75.00	9.00	9.00	9.00	9.00	9.00	9.00	9.00	9.00
7	Index_Utilities	11.00	35.00	5.00	3.00	5.00	3.00	5.00	3.00	5.00	3.00
8	Index_Healthcare	10.00	185.00	15.00	27.00	15.00	27.00	15.00	27.00	15.00	27.00
9	Index_Conglomerates	6.00	4.00	1.00		1.00		1.00		1.00	

Figure 14. Match Matrix Unique Word Pairs by Theme

[2,0.15]SEES ENERGY(101, Popularity)
[2,0.14]POISED ENERGY(101, Popularity)
[2,0.40]UPCOMING ENERGY(101, Popularity)
[3,0.16]CAP OIL(101, Popularity)
[4,0.36]PRODUCTS PARTNERS_LP(101, Popularity)
[4,0.50]UPDATES DRILLING(101, Popularity)
[2,0.25]UPDATES RESOURCES(101, Popularity)
[3,1.00]MEMORIAL PRODUCTION(101, Popularity)
[2,0.29]MARCELLUS PRODUCTION(101, Popularity)
[3,0.25]SPECIAL STOCK(101, Popularity)
[2,1.00]PRODUCT PRICES(101, Popularity)
[2,0.67]APPROVES DIVIDEND(101, Popularity)

Figure 15. Drill-Down List for the Unique Word Pairs in Theme 101 and Index_BasicMaterial

LLA Search Results

"SEES ENERGY" returned 487 results (**vis**)

12-7-2011_NS.html
REFILE NUSTAR ENERGY SEES GROWTH ACROSS ITS BUSINESSES
http://dise4.ern.nps.edu:8080/CLA/publish/Index_BasicMaterials/12-7-2011_NS.html
(20004.00 ~ 4.00,nustar energy,refile nustar,sees growth,sees energy,)

12-7-2011_ALXN.html
GULFPORT ENERGY IS A TOP PICK IN MID CAP GROWTH INVESTING DUE TO THEIR GOOD PRODUCTION GROWTH AND VERY STRONG EXCESS CASH FLOW MONEY MANAGER AT M AND I INVESTMENT MANAGEMENT REVEALS ALL HIS PICKS IN THIS EXCLUSIVE INTERVIEW
http://dise4.ern.nps.edu:8080/CLA/publish/Index_Healthcare/12-7-2011_ALXN.html
(10016.00 ~ 16.00,gulfport energy,production growth,exclusive interview,mid cap,due investing,flow money,flow cash,excess cash,investment management,)

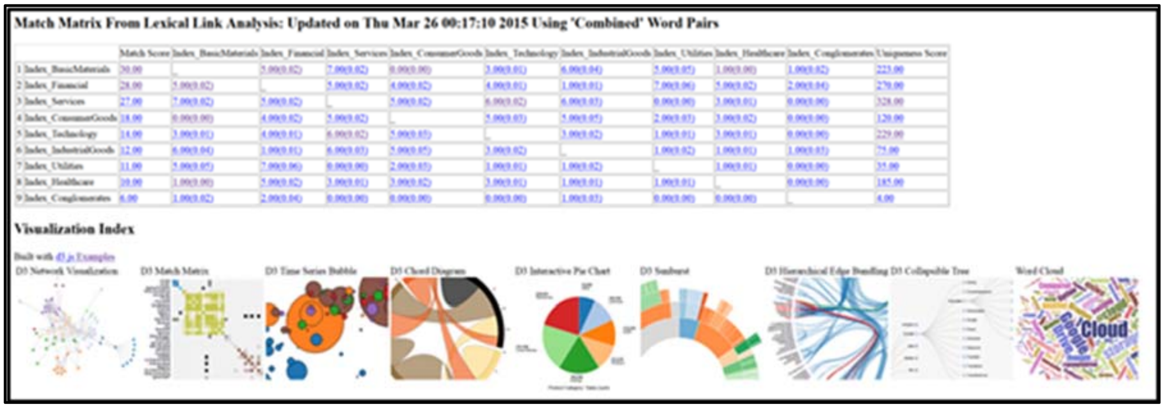
12-7-2011_ADS.html
GULFPORT ENERGY IS A TOP PICK IN MID CAP GROWTH INVESTING DUE TO THEIR GOOD PRODUCTION GROWTH AND VERY STRONG EXCESS CASH FLOW MONEY MANAGER AT M AND I INVESTMENT MANAGEMENT REVEALS ALL HIS PICKS IN THIS EXCLUSIVE INTERVIEW
http://dise4.ern.nps.edu:8080/CLA/publish/Index_Services/12-7-2011_ADS.html
(10016.00 ~ 16.00,gulfport energy,excess cash,flow money,production growth,mid cap,interview exclusive,investment management,cap growth,investing due,)

12-7-2011_GPOR.html
GULFPORT ENERGY IS A TOP PICK IN MID CAP GROWTH INVESTING DUE TO THEIR GOOD PRODUCTION GROWTH AND VERY STRONG EXCESS CASH FLOW MONEY MANAGER AT M AND I INVESTMENT MANAGEMENT REVEALS ALL HIS PICKS IN THIS EXCLUSIVE INTERVIEW
http://dise4.ern.nps.edu:8080/CLA/publish/Index_BasicMaterials/12-7-2011_GPOR.html
(10016.00 ~ 16.00,gulfport energy,flow money,mid cap,investing due,reveals management,excess cash,interview exclusive,cap growth,excess strong,)

Figure 16. Drill-Down Search on "Sees Energy"

Match Matrix

Figure 17 shows the match matrix for comparing data sources. The column “Match Score” shows the number of matched word pairs for Index_BasicMaterials. “5.00(0.02)” shows the number (5) of matched word pairs and correlation (0.02) between Index_BasicMaterials and Index_Financial. The correlation, computed as *=5/(sqrt(30+223)*sqrt(28+270))*, is normalized using the match scores and uniqueness scores for both data sources. Clicking on the “5.00(0.02)” leads to the list of the matched word pairs for the two sources as shown in Figure 18. Clicking on “Energy Prices” or “Oil Fund” (i.e., the red boxes in Figure 18) leads to the search results of two terms respectively. The search results are sorted in a descending order of the counts of how many “popular,” “emerging,” and “anomalous” word pairs appear in the original documents. For example, some marketing applications may need listing the popular terms, and business intelligence applications may need listing anomalous terms as shown in Figure 19 (a) and (b) respectively. Clicking the “vis” link in Figure 19 (a) and (b) lead to the corresponding themes to which these word pairs (e.g., “energy prices” and “oil fund”) belong.



Match Matrix From Lexical Link Analysis: Updated on Thu Mar 26 00:17:10 2015 Using 'Combined' Word Pairs

	Match Score	Index_BasicMaterials	Index_Financial	Index_Services	Index_ConsumerGoods	Index_Technology	Index_IndustrialGoods	Index_Utilities	Index_Healthcare	Index_Conglomerates	Uniqueness Score
1 Index_BasicMaterials	30.00	_	5.00(0.02)	7.00(0.02)	0.00(0.00)	3.00(0.01)	6.00(0.04)	5.00(0.05)	1.00(0.00)	1.00(0.02)	223.00
2 Index_Financial	28.00	5.00(0.02)	_	5.00(0.02)	4.00(0.02)	4.00(0.01)	1.00(0.01)	7.00(0.06)	5.00(0.02)	2.00(0.04)	270.00
3 Index_Services	27.00	7.00(0.02)	5.00(0.02)	_	5.00(0.02)	6.00(0.02)	6.00(0.03)	0.00(0.00)	3.00(0.01)	0.00(0.00)	328.00
4 Index_ConsumerGoods	18.00	0.00(0.00)	4.00(0.02)	5.00(0.02)	_	5.00(0.03)	5.00(0.05)	2.00(0.03)	3.00(0.02)	0.00(0.00)	120.00
5 Index_Technology	14.00	3.00(0.01)	4.00(0.01)	6.00(0.02)	5.00(0.03)	_	3.00(0.02)	1.00(0.01)	3.00(0.01)	0.00(0.00)	229.00
6 Index_IndustrialGoods	12.00	6.00(0.04)	1.00(0.01)	6.00(0.03)	5.00(0.05)	3.00(0.02)	_	1.00(0.02)	1.00(0.01)	1.00(0.03)	75.00
7 Index_Utilities	11.00	5.00(0.05)	7.00(0.06)	0.00(0.00)	2.00(0.03)	1.00(0.01)	1.00(0.02)	_	1.00(0.01)	0.00(0.00)	35.00
8 Index_Healthcare	10.00	1.00(0.00)	5.00(0.02)	3.00(0.01)	3.00(0.02)	3.00(0.01)	1.00(0.01)	1.00(0.01)	_	0.00(0.00)	185.00
9 Index_Conglomerates	6.00	1.00(0.02)	2.00(0.04)	0.00(0.00)	0.00(0.00)	0.00(0.00)	1.00(0.03)	0.00(0.00)	0.00(0.00)	_	4.00

Visualization Index

Built with d3.js Examples

Figure 17. Match Matrix and Visualization List

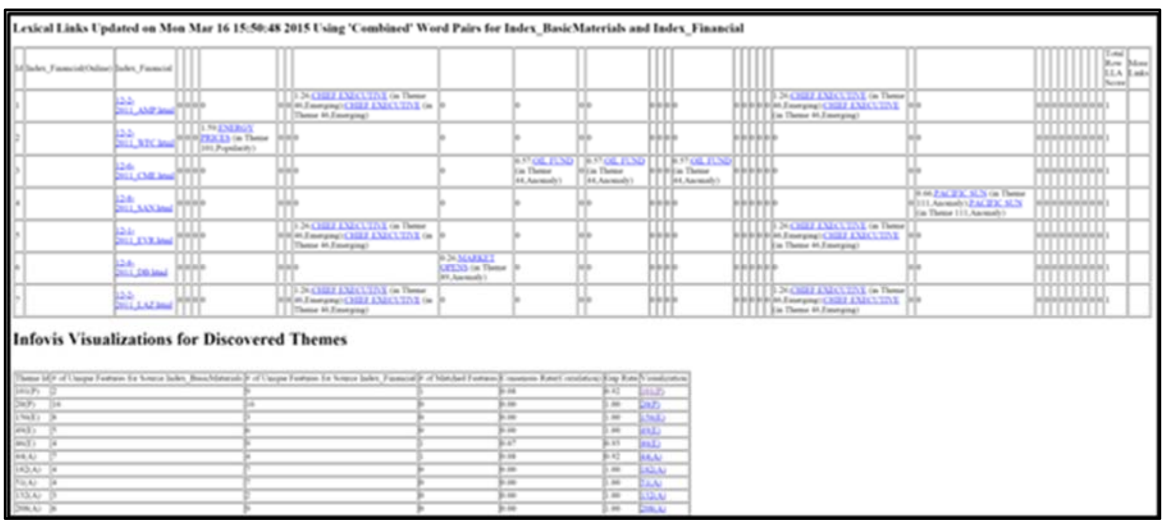


Figure 18. Drill-Down (e.g., Correlation Between Index_BasicMaterials and Index_Financial)

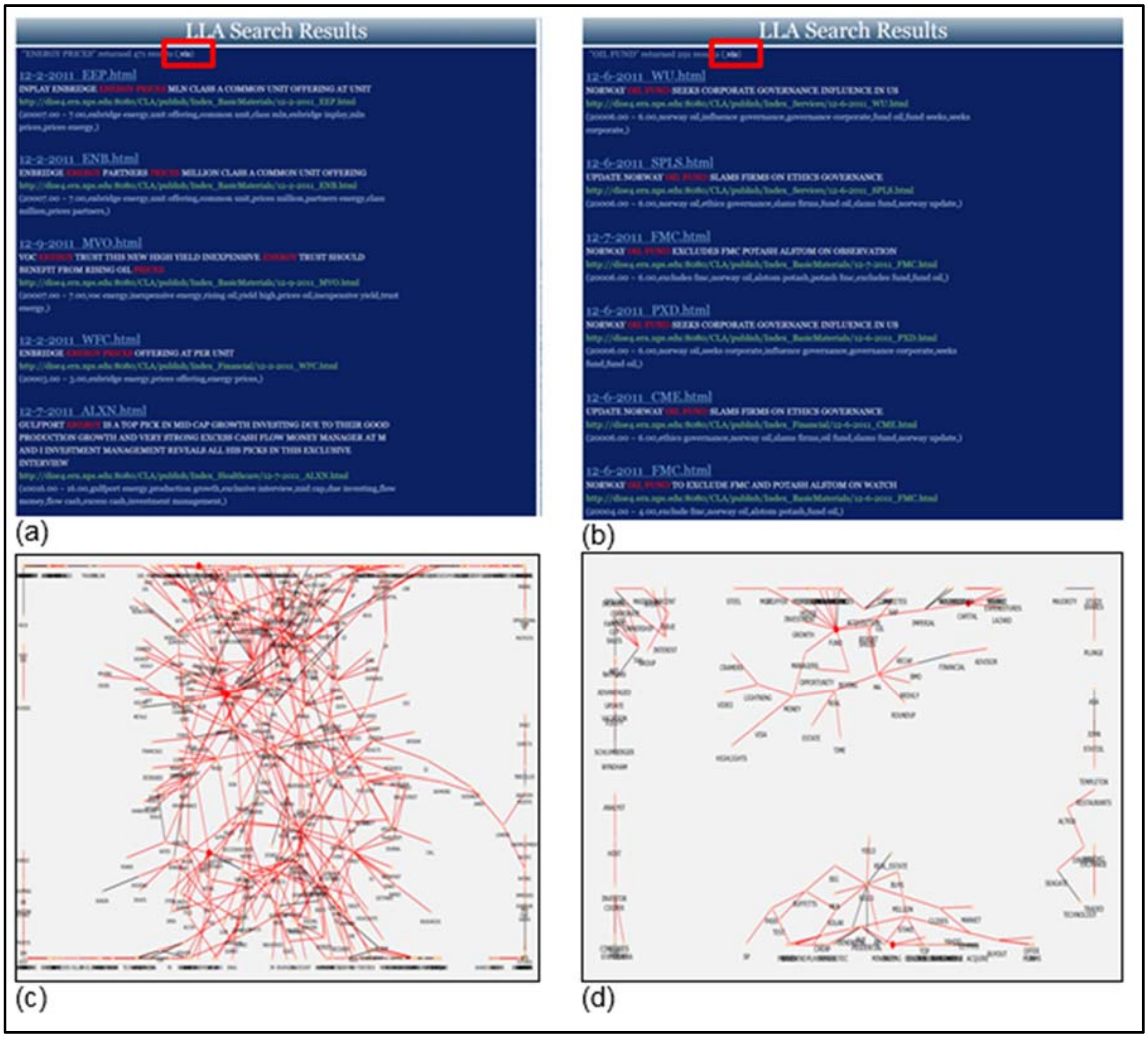


Figure 19. Drill-Down Options From Figure 10

Figure 17 also includes a list of D3 visualizations implemented. Figure 20 shows a D3 network visualization for all the data sources; their connections among the nodes are computed based on the correlations from the lexical links in Figure 17. The node connections represent all the correlations: thicker (thinner) connections indicate higher (lower) correlations. The clusters are generated based on the correlations. Figure 21 shows a D3 correlation matrix view of all the data sources. Figure 22 shows a D3 time-series bubble chart, which depicts the changing of the themes over time.

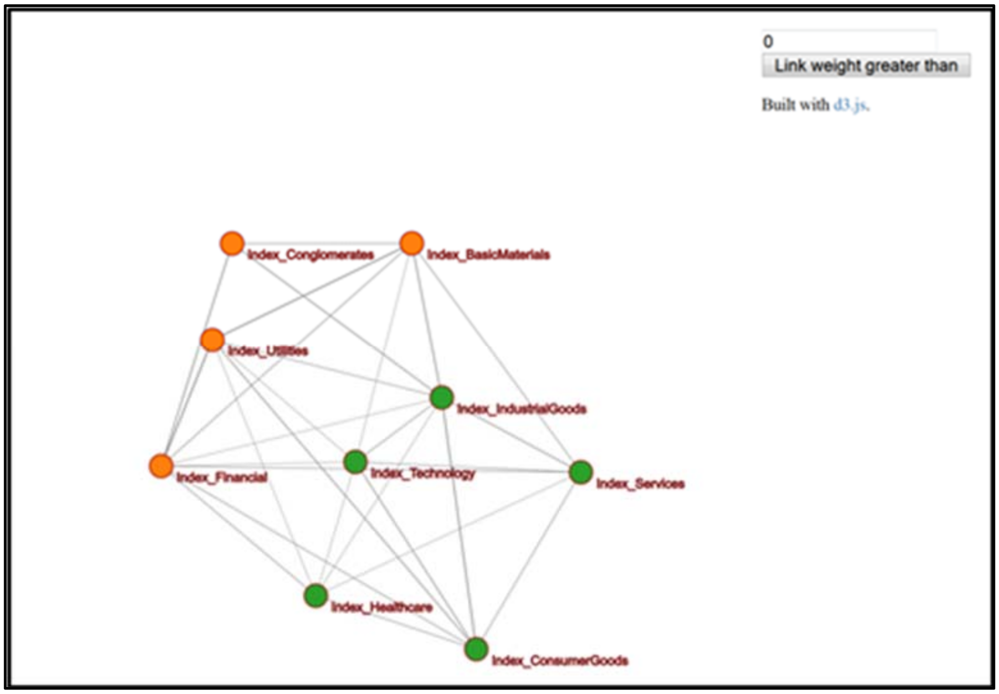


Figure 20. D3 Network

Figure 21. D3 Correlation Matrix

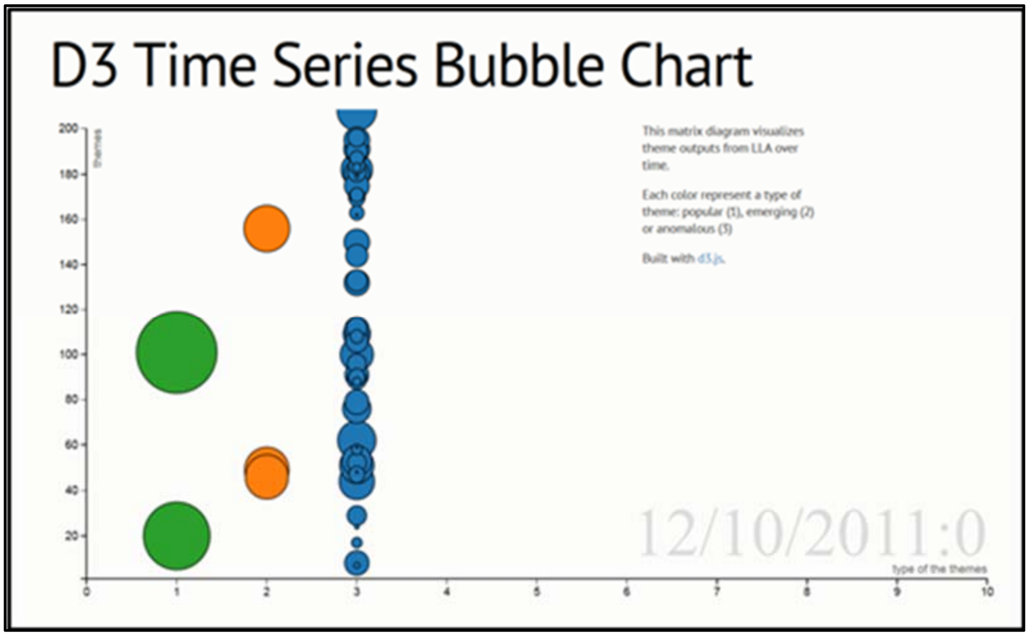


Figure 22. D3 Time-Series Bubble Chart

Task 2

We are also exploring how to use LLA jointly with other business intelligence tools, especially Big Data Architecture and Analytics (BDAA) tools:

- Deep learning, machine vision, large-scale object identification across heterogeneous data sources. One important trend in Big Data is Deep Learning, including unsupervised machine learning techniques (e.g., neural networks) for recognizing objects of interest from Big Data [9], for instance, sparse coding (Olshausen & Field, 1996) and self-taught learning (Raina et al., 2007). The self-taught learning approximates the input for unlabeled objects as a succinct, higher-level feature representation of sparse linear combination of the bases. It uses the Expectation and Maximization (EM) method to iteratively learn coefficients and bases. Deep Learning links machine vision and text analysis smartly. For example, text analysis Latent *Dirichlet* Analysis (LDA) is a sparse coding where a bag of words is used as the sparsely coded features for text (Olshausen & Field, 1996). Our methods Lexical Link Analysis (LLA), System-Self-Awareness (SSA), and Collaborative Learning Agents (CLA) can be viewed as unsupervised learning or Deep Learning for pattern recognition, anomaly detection, and data fusion.

A recursive data fusion methodology leveraging LLA, SSA, and CLA can be employed as follows:

- An agent *j* represents a sensor, operates on its own like a decentralized data fusion; however, it does not communicate with all other sensors but only with the ones that are its peers. A peer list can be specified by the agent.
- An agent *j* includes a learning engine CLA that collects, analyzes from its domain specific data knowledge base *b(t,j)*—for example, *b(t,j)* may represent the statistics for bi-gram feature pairs (word pairs) computed from LLA.
- An agent *j* also includes a fusion engine SSA with two algorithms SSA1 and SSA2 that can be customized externally. SSA1 integrates the local

knowledge base *b(t,j)* to the total knowledge base *B(t,j)* that can be passed along to its peers and used globally in the recursion in Figure 6. SSA2 assesses the total value of the agent j by separating the total knowledge base into the categories of patterns, emerging and anomalous themes based on the total knowledge base *B(t,j)* and generates a total value *V(t,j) as follows*:

- *Step 1: B(t,j) = SSA1(B(t-1, p(j)), b(t,j));*
- *Step 2: V(t,j) = SSA2(B(t,j))*
  where *p(j)* represents the peer list of agent *j*.

- The total value *V(t,j)* is used in the global sorting and ranking of relevant information. In this recursive data fusion, the knowledge bases and total values are completely data-driven and automatically discovered from the data. Each agent has the exact same code of LLA, SSA, and CLA, yet has its own data apart from other agents. This agent work has the advantages of both decentralized and distributed data fusion. It performs learning and fusion simultaneously and in parallel. Meanwhile, it categorizes the patterns and anomalous information.
- Spark (2015): Map/Reduce is an analytic programming paradigm for Big Data. It consists of two tasks: (1) the "Map" task, where an input dataset is converted into key/value pairs; and (2) the "Reduce" task, where outputs of the "Map" task are combined to a reduced key-value pairs. Apache Spark could replace Map/Reduce for its speed and in-memory computation.
- Bayesian Networks with R and Hadoop (Mendelevitch, 2015): It is a data-driven learning of conditional probability or structure learning. It is a supervised learning method but best for Big Data with low dimensions. It is an approximate inference good for Big Data and Hadoop implementation.

We have also met the acquisition professionals and discussed how BDAA can be applied to the DoD acquisition process; the following is a summary of the findings:

1. In the current acquisition process, a small delay or anomaly in a contract negotiation process can have a huge impact in its performance and can therefore cost the government a lot of money downstream.
2. It will be very useful to apply BDAA such as LLA for pattern recognition and anomaly detection for these kind of problems and make early warnings and predictions to prevent the downstream risks.
3. The Big Acquisition Data might include programs' cost/EUM, SAR, DIMIR, tech data, people data from DMBC, even outside economic environment data if the access is possible.
4. The causes of the deviations from the normal behaviors for the programs/contracts might be modeled using physics (e.g., fluid dynamics theories).
5. LLA's network perspectives, social plays among the nodes and the System Self-Awareness (SSA) theory may be used to lay out the academic vigor for the business processes, for example, answering the following questions:

- Are some nodes drawn towards some other nodes because the other nodes are more powerful?
- Is the preferential attachment growth pattern or expertise growth pattern can be used here?

- How are the forces of the nodes modeled and mapped into the social network settings and actual business processes?

Conclusion

In this paper, we show improved LLA analysis reports and visualizations generated automatically using multiple categories of data sources. These reports and visualizations reveal that there are data correlations and gaps. LLA is able to discover in detail where the gaps and inconsistencies of the data across multiple data sources reside, which, in turn, can lead to the identification of future specific and productive directions for further examination regarding why gaps occur and where they exist. It is a data-driven method for pattern recognition, anomaly detection, and data fusion. It shares indexes, not data, feasible for parallel and distributed processing, adaptive to Big Data Architecture and Analytics and needs for Big Acquisition Data.

References

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

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. In *Proceedings of the International Joint Conference on Knowledge Discovery, Knowledge Engineering, and Knowledge Management (IC3K),* Madeira, Portugal.

Girvan, M., & Newman, M. E. J. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences, USA, 99*(12), 7821–7826.

Mendelevitch, O. (2015). Bayesian networks with R and Hadoop. Retrieved from http://www.slideshare.net/ofermend/bayesian-networks-with-r-and-hadoop

Olshausen, B., & Field, D. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*.

Quantum Intelligence (QI). (2009). Collaborative learning agents (CLA). Retrieved from http://www.quantumii.com/qi/cla.html

Raina, R., Battle, A., Lee, H., Packer, B., & Ng, A. Y. (2007). Self-taught learning: Transfer learning from unlabeled data. In *Proceedings of the 24th International Conference on Machine Learning (ICML).*

Spark. (2015). Spark: Lightning-fast cluster computing. Retrieved from http://spark.apache.org/

Stanford Natural Language Processing Group (SNLPG). (2015). Retrieved from http://nlp.stanford.edu/software/lex-parser.shtml

Zhao, Y., Gallup, S. P., & MacKinnon, D. J. (2010). *Towards real-time program awareness via lexical link analysis* (NPS-AM-10-174). Retrieved from Naval Postgraduate School, Acquisition Research Program website: http://www.acquisitionresearch.net/files/FY2010/NPS-AM-174.pdf

Zhao, Y., Gallup, S. P., & MacKinnon, D. J. (2011a). *A web service implementation for large-scale automation, visualization, and real-time program-awareness via lexical link analysis* (NPS-AM-11-186). Retrieved from Naval Postgraduate School, Acquisition Research Program website: http://www.acquisitionresearch.net/files/FY2011/NPS-AM-11-186.pdf

Zhao, Y., Gallup, S. P., & MacKinnon, D. J. (2011b). Lexical Link Analysis for the Haiti earthquake relief operation using open data sources. In *Proceedings of the Sixth International Command and Control, Research and Technology Symposium* (ICCRTS), Québec City, Canada, June 21–23, 2011. Retrieved from http://www.dodccrp.org/events/16th_iccrts_2011/papers/164.pdf

Zhao, Y., Gallup, S. P., & MacKinnon, D. J. (2012a). *Applications of Lexical Link Analysis web service for large-scale automation, validation, discovery, visualization and real-time program-awareness.* (NPS-AM-12-205). Retrieved from Naval Postgraduate School, Acquisition Research Program website: http://www.acquisitionresearch.net/files/FY2012/NPS-AM-12-205.pdf

Zhao, Y., Gallup, S. P., & MacKinnon, D. J. (2012b). Semantic and social networks comparison for the Haiti earthquake relief operations from APAN data sources using lexical link analysis. In *Proceedings of the 17th ICCRTS, International Command and Control, Research and Technology Symposium*, Fairfax, Virginia, June 19–21, 2012. Retrieved from http://www.dodccrp.org/events/17th_iccrts_2012/post_conference/papers/082.pdf

Zhao, Y., Gallup, S. P., & MacKinnon, D. J. (2013). *Lexical Link Analysis application: Improving web service to acquisition visibility portal* (NPS-AM-13-109). Retrieved from Naval Postgraduate School, Acquisition Research Program website: http://www.acquisitionresearch.net/files/FY2013/NPS-AM-13-109.pdf

Zhao, Y., Gallup, S., & MacKinnon, D. (2014). Lexical Link Analysis application: Improving web service to acquisition visibility portal phase II. In *Proceedings of the 11th Annual Acquisition Research Symposium*, Monterey, CA, May 2014. Retrieved from http://www.acquisitionresearch.net/files/FY2014/NPS-AM-14-C11P17R03-065.pdf

Zhao, Y., MacKinnon, D. J., & Gallup, S. P. (2011). *System self-awareness and related methods for improving the use and understanding of data within DoD. Software Quality Professional, 13*(4), 19–31. Retrieved from http://www.nps.edu/Academics/Schools/GSOIS/Departments/IS/DISE/docs/improving-use-and-understanding-of-data-dod.pdf

Zhou, C., Zhao, Y., & Kotak, C. (2009). The collaborative learning agent (CLA) in Trident Warrior 08 exercise. In *Proceedings of the International Conference on Knowledge Discovery and Information Retrieval (KDIR).* INSTICC Press.

ACQUISITION RESEARCH PROGRAM
GRADUATE SCHOOL OF BUSINESS & PUBLIC POLICY
NAVAL POSTGRADUATE SCHOOL
555 DYER ROAD, INGERSOLL HALL
MONTEREY, CA 93943

www.acquisitionresearch.net