Unapparent Information Revelation for Counterterrorism: Visualizing Associations using a Hybrid Graph-based Approach

Norshuhani Zamin[1], Alan Oxley[2]

[1]*Universiti Teknologi PETRONAS (UTP), Malaysia, norshuhani@petronas.com.my*
[2]*Universiti Teknologi PETRONAS (UTP), Malaysia, alanoxley@petronas.com.my*

ABSTRACT

Unapparent Information Revelation refers to the task in the text mining of a document collection of revealing interesting information other than that which is explicitly stated. It focuses on detecting possible links between concepts across multiple text documents by generating a graph that matches the evidence trail found in the documents. A Concept Chain Graph is a statistical technique to find links in snippets of information where singularly each small piece appears to be unconnected. In relation to algorithm performance, Latent Semantic Indexing and the Contextual Network Graph are found to be comparable to the Concept Chain Graph. These aspects are explored and discussed. In this paper, a review is performed on these three similarly grounded approaches. The Concept Chain Graph is proposed as being suited to extracting interesting relations among concepts that co-occur within text collections due to its prominent ability to construct a directed graph, representing the evidence trail. It is the baseline study for our hybrid Concept Chain Graph approach.

Keywords: Concept Graph, Information Extraction.

I INTRODUCTION

Often documents that are generated by many authors who work independently contain interesting links that connect facts and hypotheses. The goal of Unapparent Information Revelation (UIR) is to automatically sift through these extensive documents, without human assistance, to expose a link or at least produce a plausible concept chain from those documents. UIR is considered as a challenge in text mining research. It works in a similar way to Information Retrieval (IR) but in reverse order. Traditionally, IR pulls the relevant documents from a huge collection of documents related to a query or set of queries but UIR derives a complex query that best represents a complex user need that is based on documents seen thus far.

The UIR solution benefits many domains. Earlier work has shown the success of research into finding relations. This has focused on the biomedical domain, examples include the work done on protein-protein interaction (Cooper, 2003, Yao et al, 2010), gene function and interaction (Hahn et al., 2003, Eom et al, 2005, Chaussabel & Sher, 2002, Jani et al., 2010), drug-disease association (Theobald et al., 2009, Chen et al., 2008, Rindflesch et al., 2000), identification of viruses used as bioweapons (Swanson et al., 2001), finding target diseases for the thalidomide drug (Weeber et al., 2003). Unfortunately, most of these works generated text-based output. In this paper, a graph-based method is introduced to give a better visualization of the interaction between entities in the domain of knowledge. This approach is designed specifically for web document representation (Schenker, 2003). In relation to the terrorism domain, graph-based approach benefits intelligence analysts to map terrorist activities and criminal intelligence by visualizing association between entities and events.

This paper contributes to the idea of combining different methods to achieve the objectives. We suggested a graph-based model as a potential solution to UIR. Hence, the term 'hybrid graph-based' refers to the use of good properties in existing graph-based research to problems they can efficiently solve.

A brief discussion is presented in this paper on the possibility of associating links between concepts across multiple documents using a graph-based approach known as the Concept Chain Graph (CCG). In addition, a review of the relative performance of CCG, Contextual Network Graph (CNG) and Latent Semantic

Indexing (LSI), in the terrorism domain, is also tabled.

II RELATED WORK

Information overload is now a reality. The overload of documents has created the potential for there to be a vast amount of valuable information buried in those texts. This has become a major motivation for the research - to explore potential algorithms and methods for discovering relevant knowledge without the user having to read everything. The research focuses on detecting plausible links between two concepts across a terrorism corpus. UIR in general is an emerging research field. It has attracted wide attention in the literature for the past few years. However, much of the work in UIR makes use of an idea generated by Swanson (Swanson, 1991). This research explains that a new interesting piece of information could be discovered if the linkage between textual records is studied. He proposed a simple linkage model of hypothesis generation – *"If A influences B and B influences C, therefore A may influence C".* This is also commonly referred to as *Swanson's ABC* model which means by linking A to B and B to C, co-occurrence network of these three concepts can be generated. It was a breakthrough for Swanson when he discovered an indirect connection between Fish Oil-Raynaud's Disease and Migraine-Magnesium (Swanson, 1991, Smalheiser & Swanson, 1996). These relationships were made using his technique, and it was two years before the medical experts found that they were real (Gordon & Lindsay, 1996). His approach benefits most research in the medical field.

Extended versions of *Swanson's ABC* model can be found in (Lindsay & Gordon, 1999) and (Weeber et al., 2003). They added NLP components to detect biomedical terms and a knowledge-based approach to identify relation based on semantic type of the terms into Swanson's model. Similarly, (Srinivasan, 2004) adopts the discovery framework proposed by Swanson to demonstrate its feasibility. The work is based on MeSH (Medical Subject Heading) terms and UMLS semantic types, which are available in the medical corpus known as MEDLINE[1]. She used MeSH terms, which are the keywords of the medical abstract to identify biomedical terms instead of the NLP components. Whilst, in text summarization, such graph-based methods allow sentences to be represented as nodes in a graph and edges connecting these nodes show the similarity between sentences as shown in (Shamsfard et. al., 2009). The similarity between two sentences is calculated using certain metrics.

Literature on text mining for counterterrorism was until recently scarce, but lately most text mining trends have been applied to the field of counterterrorism and radicalization. In this paper, our interest is focused on the CCG technique as used in counterterrorism. Relevant research includes the work of Srihari from the University of Buffalo, New York (Srihari et al, 2005a, Srihari et al, 2005b, Srihari, 2009). She proposed a UIR framework using the CCG method. This research used an existing IE framework known as InfoXtract (Srihari et al. 2003) to perform basic concept extraction processes such as tagging named entities and shallow parsing. Her proposed model is illustrated in Figure 1.

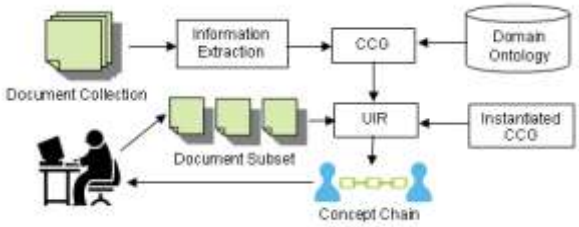


Figure 1. Model for creating and mining CCGs (Adopted from (Srihari et al. 2003))

In this paper, CCG is studied and presented based on this model. CCG formulates the concepts (nodes) and associations (edges) as a probabilistic graph. It involves two levels of a Markov model: 1) A Markov Graph representation to find the best chain of a given length connecting two concepts; 2) A Hidden Markov Model used to retrieve the best set of documents that could have generated the chain. Srihari and her team conducted experiments on 9/11 Commission Report[2] as the dataset, has successfully produced significant evidence trails given sets of concept chain. A Concept Graph for linking terrorist of the name *Atta, Ksm* and *Hazmi* is reported in (Srihari, 2009) while the

[1] http://medline.cos.com/

[2] http://www.9-11commission.gov/report/911Report.pdf

Concept Graph for *Atta*, *BinalShibh*, *Hazmi* and *Imam Anwar Aulaqui* is reported in (Srihari et al, 2005b).

Conversely, LSI and CNG seem to show great promise in UIR. LSI[3] is a patented content discovery algorithm that has produced great achievements in addressing the problems of synonymy in text collections. LSI examines the document collection as a whole. It considers documents with words in common to be semantically connected, and those with few words in common to be semantically secluded. LSI can be implemented with two kinds of algorithm: 1) Singular Value Decomposition (SVD) and 2) a probabilistic technique. "The method builds a semantic space, map each term into this space and cluster automatically according to the meaning of terms." (Farhoodi et.al., 2009). A Term Document Matrix (TDM) is a weighted lookup table of term frequency data generated by LSI for the entire document collection. Successful features of LSI are: 1) The ability to return results for queries even when an exact keyword match is not found; 2) Its support for a non-text based (mixed) and numerical domain (Landauer, 2007, Choudhury, 2002). Integrating LSI with graphing tool allows relationship information represented in link chart as described in (Bradford, 2006). Bradford performed an experiment on Salafist Group for Call and Combat (GSPC) dataset using LSI has successfully discovered how *Abu Doha*, the senior representative of GSPC is connected to other terrorist groups. However, LSI also has some drawbacks: 1) Poor scalability of the SVD in LSI is an obstacle to indexing very large document collections (Ceglowski et al., 2003); 2) For every incremental update in changing documents or increase in database, it is very costly to recomputed the SVD of the new-term-by-document matrix collections (Berry et al. 1999); 3) LSI has difficulty handling polysemy problem i.e. a word with more than one meaning (Deerwester et al. 1990) On the other hand, the graph-based Ceglowski's CNG (Ceglowski et al., 2003) appears to be an alternative to LSI. CNG does not encode any information about grammatical or hierarchical relationships between terms. It provides a less intensive way to find semantic relationships as compared to LSI. A document's term nodes are connected by edges, and the strength of each edge is determined by the frequency co-occurrence of terms across the document collection. CNG produces a network in which similar documents are closely connected and dissimilar documents are less closely connected. Apart from it being less computational intensive, CNG is found to have another advantage over LSI. CNG is very useful when working with a large collection of unstructured data. Its corpora can be extended or modified on the fly without the need to recalculate the entire index structure. Conversely, LSI requires recalculation in order to maintain its index integrity.

III UIR ARCHITECTURE

Knowledge discovery is a process that starts with literature and data related to a particular problem and attempts to look for interesting relationships that lead to potential discovery of new information. A UIR solution is expected to automate part of this process. Our UIR solution comprises of the following components: 1) Information Extraction (IE) - to identify meaningful concepts from target documents; 2) Information Retrieval (IR) - to provide co-occurrence statistics or terms; 3) A UIR Engine - to generate plausible linked concepts and interpretation. The proposed solution architecture is presented in Figure 2.

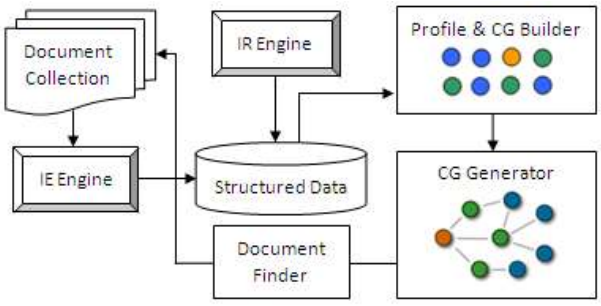


Figure 2. Proposed UIR Architecture

In this proposed system, the mining process of important terms is done by the IE component. Additionally, a user may formulate a query that will select an initial subset of target documents from a large database. The research will use free IE software such as the *RapidMiner*[4] and *Calais*[5]. An IR indexer provides corpus-wide statistics on

[3] http://www.patentstorm.us/patents/7152065/fulltext.html

[4] http://rapid-i.com/content/view/73/148/

[5] http://www.opencalais.com/about

individual words or terms. These may include concepts and their associations and semantic types. The research uses a free text indexing toolkit known as *Swish-e*[6]. Finally, the results retrieved by the IE and IR components are then processed by the UIR Engine. In this component, the Concept Selector selects prominent concepts based on their probabilistic weight. The CCG Generator generates the most likely path connecting two or more concepts. The system is equipped with a Document Finder to validate the conclusion by relating concepts and associations to underlying text snippets in documents. Without this feature, users need to validate the conclusion by looking at all actual documents. Every time the user formulates the query, a new CCG is added from a new set of documents.

IV TERRORISM DATABASES

Since September 11, 2001, the academic research on terrorism, counterterrorism and radicalization has expanded dramatically. Today, a number of publicly accessible databases comprise of current research projects. Trends as well as information on terrorist events and incidents around the world are available. These have been major sources for creating the terrorism corpora used as our document collection. Following are some of the available terrorism databases: Global Terrorism Database (GTD)[7], RAND Database of Worldwide Terrorism Incidents[8], Terrorism, Counterterrorism and Radicalization Database[9] and MIPT Terrorism Knowledge Base[10]. GTD appears to be the largest database with more than 80,000 domestic and international terrorist incidents recorded between 1970 and 2007.

V HYBRID CONCEPT CHAIN GRAPH

A CCG is a statistical and graph-theoretic approach in text mining focusing on detecting a link between two topics across unstructured texts. A CCG allows a user to query the collections to find the most meaningful evidence trails and presents the result as a bipartite graph. This technique helps to reduce the burden of having to do cumbersome modeling. The following figures illustrates the example of a concept chain connecting *Amir Abdelgani* and *Mohammad Saleh* in the corpus using the concepts *fuel* and *American wife*. The sub graph in Figure 4 is the expected CCG showing the plausible connections between prominent concepts from the textual evidence trail in Figure 3.

Mohammed Saleh, who provided **fuel** from his Yonkers gas station to make bombs, obtained legal permanent residency by marrying an **American**. Ibrahim Ilgabrowny passed messages between conspirators and obtained five fraudulent Nicaraguan passports for his cousin, El Sayyid Nosair, and his family. Nosair, convicted of conspiracy, married an **American** in 1982 and became a citizen in 1989. He was also convicted of a gun charge in the killing of Rabbi Meir Kahane in 1990. **Amir Abdelgani** picked up **fuel** and helped determine targets; he, too, was married to an **American**.

Figure 3. Text Evidence Trail (Jin et al., 2007a).

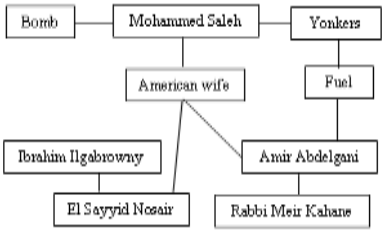


Figure 4. CCG for a Single Document

Our hybrid CCG approach consists of three main modules: A) Profile and Concept Chain Builder, B) Concept Chain Generator and C) Document Finder.

A. Profile and Concept Chain Builder

This part is adapted from the text mining algorithm in (Srinivasan, 2004). The algorithm was designed to identify interesting concepts in MEDLINE. MEDLINE is a premier bibliographic database developed by the U.S. National Library of Medicine (NLM)[11] covering the fields of medicine, nursing, dentistry, veterinary, health care systems, etc. The database was to support medical research. In our research the algorithm will be tested on a terrorism corpus. There are three steps involved in this module:

1) Building the topic profile:

[6] http://swish-e.org/
[7] http://www.start.umd.edu/gtd/
[8] http://www.rand.org/nsrd/projects/terrorism-incidents/
[9] http://www.terrorismdata.leiden.edu/
[10] http://www.mipt.org/
[11] http://www.nlm.nih.gov/

A profile is a set of prominent concepts that represent the corresponding topic. A topic profile is built by first identifying a relevant subset of documents from a text collection, next identifying characteristics (single words / phrases) from this subset, and finally, assessing their relative importance as descriptors of the topic. However, concept extraction requires an IE engine to tag named entities, common relationships associated with persons and organizations.The profiles are weighted vectors of concepts as shown below for topic T_i:

$$Profile(T_i) = \{w_{i,1}m_1, w_{i,2}m_2, \ldots, w_{i,n}m_n\} \quad (1)$$

where m_j represents a concept, $w_{i,j}$ is its weight and there are n concepts in topic T_i. Topics may be in free text format, not explicitly mentioning concepts. For example, the profile for the topic *U.S Embassy Bombing 1998* may consist of the following concepts: *Osama bin Laden*, *Egyptian Islamic Jihad*, *Al-Aqsa*, *Nairobi* and *Hamdan Khalif Alal*.

2) *Employing semantic type in a profile:*
Profiles are simply vectors of weighted concepts. 'Semantic type' is proposed as an element to differentiate concepts. These semantic types will be provided by the IR component, specifically the previously mentioned Swish-e, a free IR toolkit. A profile is a vector of concepts corresponding to topics grouped by semantic type. The concept weights are computed within the context of the semantic type:

$$Profile(T_i) = \{\{w_{i,1,1}m_{1,1}, \ldots, w_{i,n,1}m_{n,1}\}, \{w_{i,1,2}m_{1,2}, \ldots, w_{i,n,2}m_{n,2}\}..\} \quad (2)$$

where $m_{x,y}$ represents the concept m_y that belongs to the semantic type x and $w_{i,x,y}$ is the computed weight for $m_{x,y}$.

3) *Computing and normalizing concept weights:*
The Term Frequency–Inverse Document Frequency (TF-IDF) algorithm (Jones, 1972) is used to compute and normalize a concept's weight. The main idea of the TF-IDF weighting scheme is that concepts should be weighted according to the collection frequency, so that matches of less frequent, more specific, concepts are of greater value than matches of frequent concepts. This weight is a statistical measure to indicate how important a concept (word / phrase) is to a document in a collection (corpus):

$$w_{i,x,y} = \frac{v_{i,x,y}}{highest(v_{i,x,l})} \quad (3)$$

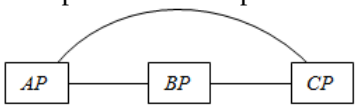
where $l = 1, \cdots, r$ and $v_{i,x,y} = n_{i,x,y} * log(N/n_{x,y})$. Here N is the number of documents in the database, $n_{x,y}$ is the number of documents in which $m_{x,y}$ occurs and $n_{i,x,y}$ is the number of retrieved documents for T_i in which $m_{x,y}$ occurs. Normalization by $highest(v_{i,x,l})$, the highest value for $v_{i,x,y}$ observed for the concepts with semantic type x, yields weights that are between 0 and 1 within each semantic type. Note that there are r terms in the domain for the semantic type x.

B. Concept Chain Generator

A Concept Chain is considered a graph G (N, E) representing a document collection (Jin et al., 2007a, Jin et al., 2007b). It is a weighted labeled graph where N is a set of nodes; E is a collection of weighted edges. A node represents a concept in the document collection. An edge is constructed based on the proximity and co-occurrence relationships between concepts. If the two concepts co-occur within a paragraph or sentence, then there is an edge connecting them whilst a weight represents the measure of how strong the association is. In what follows a concept chain is created and ranked according to the weight of the corresponding selected concept. The result is a potential conceptual connection at different levels. The underlying procedure is as follows:

1) The algorithm takes two inputs - two concepts of interest. Assume these to be *A* and *C*. Build the *A* and *C* profiles, refer to these as *AP* and *CP* respectively. Associated with the link, the user can specify semantic types that are common to *AP* and *CP*.

2) According to the semantic types for intermediate concepts specified by the user, compute a *B* profile (*BP*) composed of terms in common with *AP* and *CP*, within the specified semantic types. The weight of a concept in *BP* is the sum of its weights in *AP* and *CP*. Concepts are retained and ranked by an estimated potential for each specified semantic type. This is the first level of intermediate potential concepts.



Weight $w_{A,B}$ can be calculated as the co-occurrence frequency of concept *A* and *B* within the similar paragraph / sentence. To measure the association degree, a formula analogous to Dice Coefficient[12] is used:

$$w_{A,B} = log\ (1 + F\ (A,B)) \quad (4)$$

where $F\ (A, B) = 2 * N_{A,B} / (N_A + N_B)$.
N_A is the co-occurrence frequency of concept *A* in the document collection. N_B is the co-occurrence frequency of concept *B* in the document collection $N_{A,B}$ is the co-occurrence frequency of concept *A* and *B* within the paragraph / sentence. Based on this model each document can be represented as follows:

$$CL = [C_1, C_2, \cdots, C_n] \quad (5)$$

$$M = [a_{ij}] \quad (6)$$

where *CL* is the concept list, C_i is the i^{th} concept in the concept dictionary, *M* is the strictly upper triangular matrix among concepts and a_{ij} is the association strength between concept C_i and C_j $(1 \le i < j \le n)$.

3) In the next level onwards, expand the concept chain using the created *BP* profile together with the topics to build additional levels of intermediate concept lists. Eventually, find the best path from *A* to *C* based on concepts ranked by their weights within the specified semantic types. "A potential conceptual connection between *A* and *C* is a path starting from topic *A*, proceeding through sequential levels of intermediate concepts until reaching the ending topic *C*" (Jin et al., 2007a).

C. **Document Finder**

QuExt (Matos et al. 2010) is a document retrieval system that searches for the most relevant results from the MEDLINE literature given a list of genes. The *Document Finder* module helps to retrieve the best set of documents that could have generated the chain. We proposed the document weighting scheme in QuExt to meet our needs. The results from the *Concept Chain Generator* module are assembled and documents are re-ranked in terms of the defined weights for each concept. The final score for document *i* is obtained as a weighted sum of the concept-based scores:

$$Score_i = \sum_{j=1}^{n} W_j * S_{ij} \quad (7)$$

where W_j is the weight attributed to the concept *j* and S_{ij} represents the score for document *i* in terms of the j^{th} concept type. The document with the highest score is considered to be the most relevant to the CCG.

VI EMPIRICAL COMPARISON

This section provides direct comparisons between CCG, LSI and CNG giving some insights on the approaches performance against terrorism data. The comparison is based on a literature search and our own observation and investigates several aspects of each approach including programming effort, run-time efficiency, memory consumption, robustness and reliability. The theoretical advantages of those approaches have been reviewed and results are shown in Table 1.

Table 1. Investigation Summary

Approach (Main Research)	Programming Complexity	Run-time Efficiency	Memory Consumption	Robustness	Test Data / Number of Articles	Method Type
CCG (Srihari	H	M	H (Matrix)	H	9/11 Report / NA	Stochastic

[12] http://en.wikipedia.org/wiki/Dice's_coefficient

et al, 2005b)			L (List)			
LSI (Bradford, 2006)	H	M	H	M	Terrorism Articles / 150,000	Stochastic
CNG (Ceglowski et al., 2003)	H	M	H (Matrix) L (List)	M	American Civil War Articles / NA	Stochastic

(**Note**: H = High, M = Medium, L = Low)

Although statistical analysis is available to proof the performance for some research method, they are not of interest to be quantitatively reported here. In this study, the methods are best comparable within the specific aspect of the problem i.e. graph-based visualization. However, this basic evaluation review has motivated the author to experiment the potential of methods using local terrorism data. In addition, it is important to explore the possibility to improve the existing research (Srihari et al, 2005b). The implemented algorithm would be a vehicle to precisely understand it with a view to suggesting a refined algorithm.

VII CONCLUSIONS

This paper focuses on a hybrid technique to detect hidden links between two topics of interest. The technique is referred to as UIR. The solution architecture for generating a CCG is described. The generated CCG is a probabilistic network with the nodes representing concepts and edges between them representing associations. The task of categorizing documents based on their conceptual similarities, has demonstrated the superiority of the graph theoretic approaches, particularly the CCG, over other approaches for extracting semantic information from documents. However, we have identified that noisy texts and the size of documents are the possible challenges. Thus, these properties will be considered in the experimental set up to further prove the hypothesis.

The development of a hybrid CCG is currently underway. Later, the CCG will be evaluated against CNG, LSI and human-classified collections to obtain empirical estimates of search quality. Currently, the local intelligence analysts manually construct domain models that can be matched against data collections to look for a scenario of interest such as identifying an actor's importance roles among a specific group of entities in a terrorist network. It is hoped that this research contributes through the development of automatic methods to generate useful hypothesis with little human intervention.

VII ACKNOWLEDGMENT

We would like to thank the anonymous reviewer of the conference for the valuable comments and suggestions to improve the paper.

REFERENCES

A. Choudhury, Y.S. Ong and A.J. Kean. (2002) Extracting Latent Structures in Numerical Classification: An Investigation using Two Factor Models. *In Proceedings of 9th International Conference on Neural Information Processing.* IEEE. Institute of Electrical and Electronics Engineers, 4:1842-1846.

A. Schenker. (2003). Graph-Theoretic Techniques for Web Content Mining. *Ph.D Thesis*, University of South Florida, Florida, USA.

D. Chaussable and A. Sher. (2002). Mining Microarray Expression Data by Literature Profiling. *Journal of Genome Biology*, 3(10): 1-16.

D. R. Swanson, N. R. Smalheiser and V. I. Torvik. (2006). Ranking Indirect Connections in Literature-Based Discovery: Swanson, D.R. (1991). Complementary Structures in Disjoint Science Literatures. *In Proceedings of the 14th ACM SIGIR*, Chicago, Illinois,USA, DOI:10.1145/122860.122889.

D.R. Swanson, N.R. Smalheiser and A. Bookstein, (2001). Information Discovery from Complementary Literatures: Categorizing Viruses as Potential Weapons. *Journal of the American Society for Information Science*, 52(10):797 -812.

E.S. Chen, G. Hripcsak, H. Xu, M. Markatou and C. Friedman. (2008). Automated Acquisition of Disease–Drug Knowledge from Biomedical and Clinical Documents: An Initial Study. *Journal of the American Medical Informatics Association*, 15(1):87-98.

J.H. Eom and B.T. Zhang. (2005). Extraction of Gene/Protein Interaction from Text Documents with Relation Kernel, *Knowledge-Based Intelligent Information and Engineering Systems*, Springer-Verlag Berlin Heidelberg, LNCS 3682:936-942.

J.W. Cooper. (2003). An Evaluation of Unnamed Relations Computation for Discovery of Protein-Protein Interactions. *In Proceedings of the SIGIR: Workshop on Text Analysis and Search for Bioinformatics*, Canada.

K.S. Jones.(1972). A Statistical Interpretation of Term Specificity and its Application in Retrieval. *Journal of Documentation*, 28(1):11-21.

L. W. Hahn, M. D. Ritchie and J. H. Moore. (2003). Multifactor Dimensionality Reduction Software for Detecting Gene-gene and Gene-Environment Interactions. *Journal of Bioinformatics*, 19(3):376-382.

L. Yao, C. Sun,. and Wang, X. (2010). Multi-class Relationship Extraction from Biomedical Literature using Maximum Entropy. *In the Proceedings of the 6th International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, DOI: 10.1109.

M. Berry, Z. Drmac and E. Jessu. (1999). Matrices, Vector Spaces, and Information Retrieval. *SIAM Review*, 41(2): 335-362.

M. Ceglowski, A. Coburn and J. Cuadrado. (2003). Semantic Search of Unstructured Data using Contextual Network Graphs. *Preliminary White Paper*, National Institute for Technology and Liberal Education, Middlebury College, Middlebury, USA.

M. Farhoodi, M. Mahmoudi, A.M.Z. Bidoki, A Yari and M. Azadnia.(2009). Query Expansion Using Persian Ontology Derived from Wikipedia. *World Applied Sciences Journal*, 7 (4): 410-417.

M. Shamsfard, T. Akhavan and M.E. Joorabchi. (2009). Persian Document Summarization by Parsumist, *World Applied Sciences Journal 7* (Special Issue of Computer & IT): 199-205.

M. Theobald, N. Shah and J. Shrager. (2009). Extraction of Conditional Probabilities of the Relationships between Drugs, Diseases, and Genes from PubMed Guided by Relationships in PharmGKB, *AMIA Summit on Translational Bioinformatics*, 124-128.

M. Weeber, R. Vos, H. Klein, L.T.W. de Jong-Van den Berg, A. Aronson and G. Molema. (2003). Generating Hypotheses by Discovering Implicit Association in the Literature: A Case Report for New Potential Therapeutic uses for Thalidomide. *Journal of the American Medical Informatics Associations*, 10(3). 252-259.

P. Srinivasan. (2004). Text Mining: Generating Hypotheses from MEDLINE. *Journal of American Society for Information Science and Technology*, 55(5): 396- 413.

R. Bradford.(2006). Application of Latent Semantic Indexing in Generating Graphs of Terrorist Networks, *In Proceedings of the IEEE International Conference on Intelligence and Security Informatics*. Springer-Verlag Berlin Heidelberg, LNCS 4011: 215-229.

R.K. Lindsay and, M.D. Gordon. (1999). Literature-based Discovery by Lexical Statistics. *Journal of the American Society for Information Science*, 50(7):574-587.

R.K. Srihari, , S. Lamkhede and A. Bhasin. (2005a). Unapparent Information Revelation: A Concept Chain Graph Approach. *In Proceedings of the ACM Conference on Information and Knowledge Management*, Bremen, Germany.

R.K. Srihari, S. Lamkhede, A. Bhasin and W. Dai. (2005b). Contextual Information Retrieval using Concept Chain Graphs, *In Proceedings of the International Workshop on Context-Based Information Retrieval*, Paris, France.

R.K. Srihari, W. Li, C. Niu, and T. Cornell. (2003). InfoXtract: A Customizable Intermediate Level Information Extraction Engine. *In Proceedings of the NAACL Worksyop of Software Engineering and Architecture of Language Technology System*, Edmonton, Canada.

R.K. Srihari. (2009). Unapparent Information Revelation: Text Mining for Counter-terrorism. *Computational Methods for Counterterrorism*, Springer-Verlag Berlin Heidelberg, LNCS DOI: 10.1007/978-3-642-01141-2_5.

S. Deerwester, S. Dumais, G. Furnas, T. Landauer and R. Harshman. (1990). Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science*, 391-407.

S. Matos , J.P. Arrais, J.M. Rodrigues, and J.L. Oliveira. (2010). Concept-based Query Expansion for Retrieving Gene Related Publications From MEDLINE. *Journal of BMC Bioinformatics*, 11(1):212.

S.D. Jani, G.L. Argraves, J.L. Barth and W.S. Argraves. (2010). GeneMesh: A Web-based Microarray Analysis Tool for Relating Differentially Expressed Genes to MeSH Terms. *Journal of BMC Bioinformatics* DOI:10.1186/1471-2105-11-166.

T. Landauer, D. Laham and M. Derr. (2004). From Paragraph to Graph: Latent Semantic Analysis for Information Visualization, *In Proceedings of the National Academy of Science*, 101:5214–5219.

T. Rindflesch, L. Tanabe, J.N. Weinstein and L. Hunter. (2000). Edgar: Extraction of Drugs, Genes and Relations in the Biomedical Literature. *In Proceedings of the Pacific Symposium on Biocomputing*, Hawaii, USA, 5:514-525.

W. Jin, R.K. Srihari and X. Wu. (2007). Mining Concept Associations for Knowledge Discovery through Concept Chain Queries, *In Proceedings of the Pacific Asia Conference on Knowledge Discovery and Data Mining*, pp. 555-562.

W. Jin, R.K. Srihari, H.H. Ho and X. Wu. (2007b). Improving Knowledge Discovery in Document Collections through Combining Text Retrieval and Link Analysis Techniques. *In Proceedings of the 7th IEEE International Conference on Data Mining*, Omaha,Nevada,USA.