Knowledge Fusion in Academic Networks

Horea Adrian Greblă
Faculty of Mathematics and Computer Science,
"Babeş-Bolyai" University of Cluj-Napoca
Mihail Kogalniceanu, 1, Cluj-Napoca, 400084, Romania,
horea@cs.ubbcluj.ro

Călin Ovidiu Cenan
Technical University of Cluj-Napoca
Memorandumului, 28, Cluj-Napoca, 400114, Romania
calin.cenan@cs.utcluj.ro

Liana Stanca
Faculty of Economics and Business Administration, "Babes-Bolyai" University of Cluj-Napoca
Teodor Mihali, 58-60, Cluj-Napoca, 400591, Romania
liana.stanca@econ.ubbcluj.ro

Abstract

Academic networks can represent a new model for learning based on knowledge fusion. In the current paper we present an approach to the academic network design that brings together expertise of academic trainers and practitioners, and opens new paths in knowledge distribution. The academic networks that we consider are mathematically modeled and they represent the foundation for the ontological approach to knowledge fusion in such a network.

Keywords: knowledge fusion, data fusion, Bayesian network based ontologies, equivalence relation, academic networks, eLearning.

1. Introduction

University degree education faces methodological transformations due to improved and modern means of knowledge presentation and sharing, transformations that turn the learning process from a restrictive into a permissive one. The term of eLearning has spread starting with the 90s in the USA, all over Europe and, since the year 2000 it has also become common in the Romanian academic training [16]. One of the main reasons for which E-Learning could not be adopted was because of the population's access to the Internet, or because individuals were not yet prepared to adopt such a model of learning. In these cases, the model was replaced by Blended Learning. According to [18], Romania takes a very honorable 4th place in the report, ahead of other countries, such as the USA. This growth in bandwidth and internet usage made eLearning adoption possible and encouraged industry players to involve in the academic learning process in various ways [4].

The emergence and development of social networks in 2003 has led to steps taken towards their integration in the academic world thus resulting in academic networks. Academic networks [17] configured as a form of a network spread at metropolitan level that links academic structures, student facilities and technological parks, etc. which can be used to develop a new learning model. This model has two players, the academic world and the non-academic, industrial one, who join efforts to develop an educational process based on quality standards. In our opinion, this new model of education is vital for the Romanian education which lacks a major percent of its practical side. Within such an educational model the courses developed and taught by the academic trainers (teachers) will be directly sustained by industry partners who work proactively in the teaching process by providing teaching laboratories, wiki logs and knowledge databases, master courses that offer practical solutions to a series of problems they are faced with. Benefits that occur based on such an approach are both for students who complete training courses and are quickly absorbed by the labor market and for employers who can eliminate vocational training of the graduates who are subject to employment after college graduation. So, in an academic network, a data (or knowledge) fusion [10] can be defined as a mix of data from industrial and academic fields, offering students

the possibility to understand themes from the theoretical and practical point of view. Information provided in the course will be composed as follows:

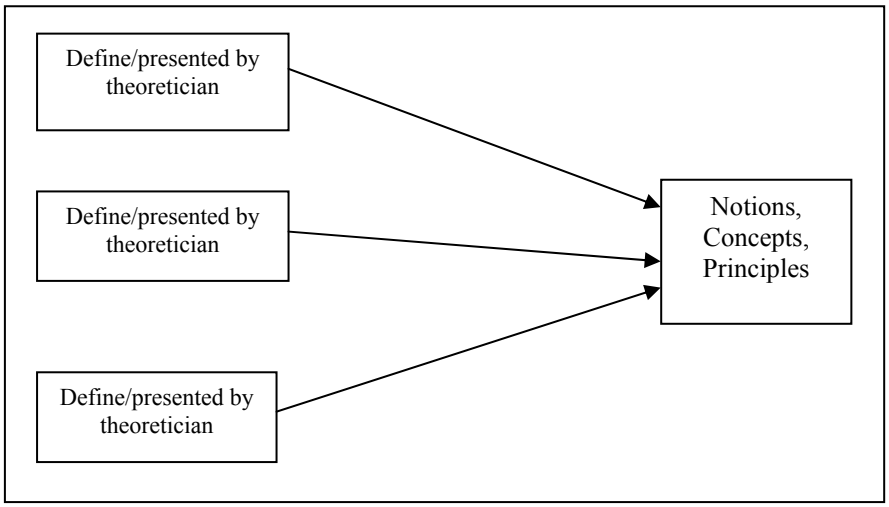

Figure 1. Data fusion in academic networks

Data fusion is generally defined [20] as the use of techniques that combine data from multiple sources and gather that information in order to achieve inferences, which will be more efficient and potentially more accurate than if they were achieved by means of a single source. This process can be seen as a set reduction technique with improved confidence.

2. Knowledge fusion and data mining

Based on our previous work in heterogeneous distributed data mining and integration [3] and based on the KRAFT Project [19], we focused our research on knowledge fusion based on anthologies.

Knowledge fusion [19] can be seen as a succession of steps that locate and extract knowledge that is stored in various forms at different locations, and on top of that data some transformations are performed so that a common representation can be applied as the foundation for problem solving. The KRAFT Project was designed as an agent-based architecture in which 3 types of agents were described: facilitator agents that are in charge of the description and location of data sources, mediator agents that are used to query data sources and fuse the knowledge collected, respectively wrappers that transform data from local format to the common format.

Our aim is to extend and adapt such an architecture in an academic network based on an ontology that is described later in this paper.

Our perception of knowledge fusion in such a network is a process that brings together specialists from various fields with the purpose of providing common solutions to different problems [9]. From Nonaka and Takeuchi‘s point of view [7], [9] knowledge fusion may result in rethinking old knowledge and work methods to obtain better results in all activity fields.

Knowledge fusion serves as a combination and transformation of various knowledge resources in order to generate new concepts / information. Knowledge fusion is a notion that stands at the intersection of the sciences of knowledge and engineering [2],[8] .

Knowledge fusion on the web, especially in academic networks, as we consider in our approach, appeared and developed following the success of collaborative software which enabled the development of applications which allow players to lead meetings and work together regardless of their geographical position. The role [15] of knowledge applications is: knowledge creation, classification, synthesis, analysis, storage, search and mapping. A common practice in the industry but increasingly used by academic networks are wikis.

Wiki [5] is a website that can be modified and revised by participants/users to knowledge applications. People using wiki can add, modify and delete information without having to possess advanced programming knowledge. Such a site gathers a synchronous and/or asynchronous discussion environment with course supports, online libraries, quizzes, exam example problems and solutions. Base principles on which the wikis are built are [5]:

1. Knowledge has a dynamic nature.
2. Knowledge obtained as a result of collective effort has a higher value than that obtained from an individual effort.

Following major changes that social values have undergone, today it is important for people to have access to knowledge [14] [15] because:

1. On the long term it offers people an advantage on a labor market characterized by a global competition,
2. In the new society based on technologies they are more valuable than natural and material resources;
3. Their owner is transformed in a group member.

Web knowledge fusion [2] [9] have two limitations:

1. The lack of knowledge presentation template makes it difficult to unify them.
2. The large amount of knowledge resources makes it difficult to manage and allocate.

So, in an academic network, "knowledge fusion" can be defined as: combining data from multiple academic and trusted non-academic sources so that it offers more levels of understanding and perspectives to all actors involved in the academic research and learning processes.

These sources can be seen as faculty libraries, department media servers where lectures, presentations, books, theses are stored, wiki systems, online scientific libraries, but also some industry partners that act proactively in the teaching process by providing hands-on labs, real world practice wiki records, master classes, etc.

Provided the above, the fusion process can be seen as a special type of data-mining.

Data-mining is concerned with extracting knowledge from databases (in this case we deal with distributed ones) using machine learning techniques. Traditionally, data-mining systems are designed to work on a single data set. However, with the increasing number of distributed databases dispersed over many machines in WANS with geographically spread locations, it is necessary to adopt new techniques to improve the overall system response [3]. The development of the Bayesian belief networks and associated algorithms made probabilistic reasoning turn into a real option for a large variety of Artificial Intelligence applications. In this paper, we address the possibility of creating a methodology study for knowledge fusion using Bayesian belief network based ontologies with domain applications in academic networks.

Later on, we develop a prototype of academic network that brings together academic training and practitioners. The academic network that we will develop will be represented as a graph which is the starting node of knowledge fusion, which will be used in the learning process like a Bayesian network.

3. A formal approach to academic networks

Academic networks can be approached using notions and results from the homogeneous binary networks theory. For this reason, we will present a concise presentation of these theories on which the academic networks are based. More details can be found in [13], [11], [12].

Let M be a set, a subset $\rho \subseteq MXM$ is a binary relation on M. Instead of the notation $(x, y) \in \rho$ we can use $(x \rho y)$. The ρ relation is called:

1. reflexive if $(x \rho x)$ for any $x \in M$
2. transitive if $x \rho y \; and \; y \rho z \Rightarrow x \rho z$
3. symmetric if $y \rho z \Rightarrow y \rho x$

A relation which is reflexive and symmetric is called a tolerance relation. A tolerance relation which is also transitive is called equivalence relation.

If $\rho \subseteq MXM$ is an equivalence relation and $x \in M$, then $\rho\langle x\rangle = \{x' \in M \mid x\rho\, x'\}$ is called an equivalence class of x by ρ. One can prove that $\rho\langle x\rangle = \rho\langle x'\rangle$ for any $x, x' \in M$ for which $x\rho\, x'$.

The set $M/\rho = \{\rho\langle x\rangle \mid x \in M\}$ is called the quotient set of M by ρ.

Theorem 1. If $\rho_i \subseteq M \times M, i \in I$ is a family of reflexive respectively transitive or symmetric relations, then $\bigcap_{i\in I} \rho_i$ is a reflexive, transitive or symmetric relation.

From this theorem it results that the intersection of a family of tolerance relations, respectively equivalence relations, is also a tolerance respectively an equivalence relation.

If $\rho \subseteq M \times M$ then the relation

$$\overline{\rho} = \bigcap \{\rho' \mid \rho' \subseteq M \times M, \rho \subseteq \rho', \rho' \textit{ reflexive}\}$$

is the smallest reflexive relation that includes ρ. The relation $\overline{\rho}$ is called the reflexive closure on ρ. The same way, one can define the transitive and the symmetric closures on ρ.

The relation

$$\bigcap \{\rho' \mid \rho' \subseteq M \times M, \rho \subseteq \rho', \rho' \textit{ is a tolerance relation}\}$$

respectively

$$\bigcap \{\rho' \mid \rho' \subseteq M \times M, \rho \subseteq \rho', \rho' \textit{ is an equivalence relation}\}$$

is the smallest tolerance relation, respectively the smallest equivalence relation that includes ρ.

A set $\{A_i \mid i \in I\}$ of subsets of M is called a partition of M if $M = \bigcup_{i\in I} A_i$ and $A_i \cap A_j = \emptyset$ *for any* $A_i \neq A_j$.

We denote by $E(M)$ the set of the equivalence relations on M and by $P(M)$ the set of the partitions of M.

Theorem 2.

1. If $\rho \in E(M)$ then M/ρ is a partition of M.
2. If $\pi \in P(M)$ then the relations ρ_π defined as:

$$x\rho_\pi y \Leftrightarrow \exists A \in \pi \textit{ such that } x, y \in A$$

is an equivalence relation on M.

The function $\varphi : E(M) \to P(M)$ is a bijection and $\varphi^{-1}(\pi) = \rho_\pi$

The evolution of the academic networks can be modeled and analysed using the mathematics theory presented above. For example

M={Mary-wiki, Jon-wiki, Victor-wiki, George-wiki, Juana-wiki}

π={{Mary-wiki, Jon-wiki},{Victor-wiki, George-wiki},{Juana-wiki}}

ρ_π={(Mary-wiki,Mary-wiki), (Jon-wiki,Jon-wiki), (Mary-wiki,Jon-wiki), (Jon-wiki,Mary-wiki), (Victor-wiki,Victor-wiki), (George-wiki,George-wiki), (Victor-wiki,George-wiki), (George-wiki,Victor-wiki), (Juana-wiki,Juana-wiki)}

M/ρ_π={{Mary-wiki, Jon-wiki },{Victor-wiki, George-wiki },{Juana-wiki }}=π

The reflexive relationship within academic networks is a relationship of a wiki-person's notes with a network of itself. Symmetry members accept each other. The transition allows that if a wiki-person's note x is linked with y and z then y can develop a relationship with z.

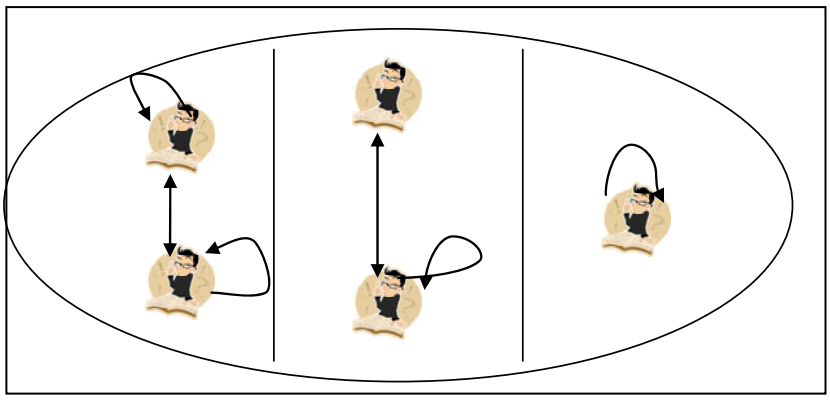

Figure 2. Equivalence classes in academic networks

4. Bayesian network based ontologies

Any scheme that tries to represent the details of a domain with a certain degree of complexity has to be expressive. In order to build an ontology, we have focused on Multi-Entity Bayesian Networks (MEBN), which are based on a language that is able to express probability distributions over interpretations of arbitrary first-order domain theories, and we have used them because of their ability to represent [21].

This language is based on first-order logic (it could be sufficient to be based on one of its subsets). A first-order theory implies truth values both for the valid sentences as well as for their negations.

A graphical probability model [22] expresses a probability distribution over a collection of related hypotheses as a graph. The graph is used to encode dependencies among the given hypotheses. The local probability distributions represent a specific numerical value for the probability information. Together, the graph and the local distributions specify a joint distribution that respects the conditional independence assertions encoded in the graph, and has marginal distributions consistent with the local distributions. Like Bayesian networks, MEBN theories [21] use directed graphs to specify joint probability distributions for a collection of related random variables. The MEBN language was built to extend basic Bayesian networks to provide first-order expressions, as well as to extend first-order logic to provide probability distributions over results of applied first-order theories.

MEBN theories [21] extend ordinary Bayesian networks. The extension refers to the structure used to represent random variables. Random variables take as arguments entities that exist in the domain of applications. Every sentence can be expressed in first-order logic and thus it can be represented as a random variable in a MEBN theory. Because of its modular and compositional nature of the language, MEBN probability distributions are specified locally over small groups of hypotheses, after that being integrated into consistent probability distributions at a global level, over sets of hypotheses.

MEBN theories can be used to express domain-specific ontologies that capture statistical regularities in a particular domain of application. MEBN theories with findings can augment statistical information with particular facts germane to a given reasoning problem. MEBN uses Bayesian learning to refine domain-specific ontologies to incorporate observed evidence.

A **distributed knowledge fusion** approach has to be constructed on top of a platform which provides all the tools necessary to build a knowledge fusion project:

- some powerful, metadata-driven ETL tools designed to bridge the gap between business and IT
- Data Mining Engine, based on a comprehensive set of tools for machine learning and data mining. It has to provide a broad suite of classification, regression, association rules and clustering algorithms that can be used to help you understand the business better and also be

exploited to improve future performance through predictive analytics. (It is a Bayesian network based tool)

- Reporting Tools, used to enrich report distribution and end user interaction. It has to enable reports that can be sent as email attachments and scheduled to run at a point in time. Report prompting using static or dynamic selections have also to be built using such a tool
- The approaches considered in this paper address both the simple data integration tool for data fusion based on transformations that are defined by users, as depicted in fig. 3 and the more complex approach of data mining using WEKA.

The following simplified Pentaho Transformation illustrates the way such a fusion can be implemented.

Data sources can be of various types, ranging from flat files, through XML streaming, Web Services and Databases. The transformations allow data and stream joins, row filtering, splitting fields, execution of scripts on certain intermediate results, data validation processes, so that the result may be more accurate and useful.

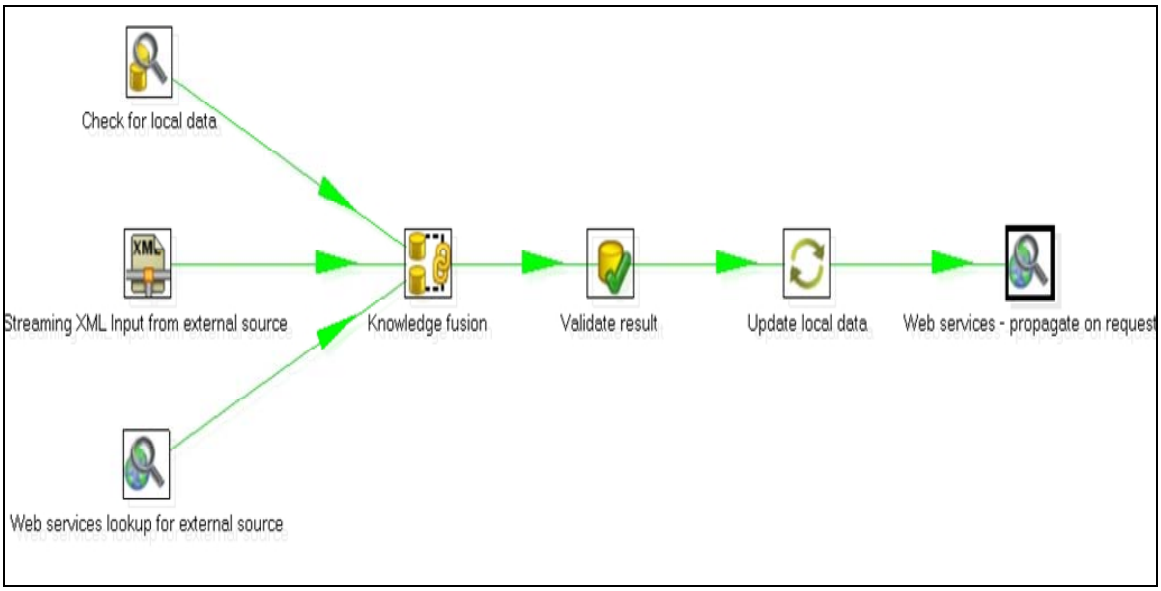

Figure 3. Simplified transformation model

The following picture (Figure 4) depicts the ontology used to model an academic network based on the mathematical formalism previously defined in this paper.

In our approach, we have considered the wiki logs for data fusion because they offer a flexible, powerful, and easy to use collaboration platform, and web application platform. A structured wiki is typically used to run a project development space, a document management system, a knowledge base, or any other groupware tool, on an intranet, extranet or the Internet. It allows users without programming skills to create web applications. Presentations are already part of the academic training process for some time, so they present proved and verified concepts. Blogs are also a good source of knowledge as every professional might want to share its knowledge but also be identified as author for some problem resolution, new theory or state of the art presentation of a certain topic.

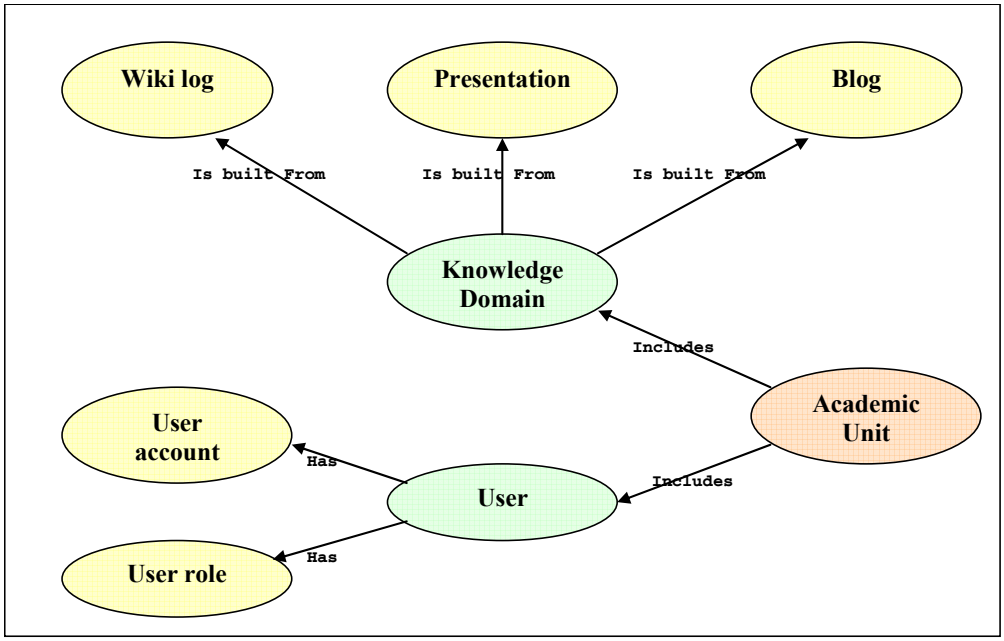

Figure 4. Academic network database model

5. Conclusions

The educational process is becoming more and more complex, involving more than traditional means of teaching. Industry can play a major role in student development. New data mining techniques and algorithms, together with some state of the art solutions found for some open problem can increase the speed at which new knowledge is discovered and accumulated. Our proposed architecture aims at bringing together the major players in student technical development (university and industry) using different knowledge interaction approaches.

References

[1] Cooper, G. F.. (1990). The computation complexity of probabilistic inference using Bayesian belief networks. *Artificial Intelligence*, 42, 393–405.
[2] Gou, J., Yang, J., & Chen, Q. (2005). Evolution and Evaluation in Knowledge Fusion System, In: Proceedings of *IWINAC 2005*, LNCS 3562 (pp. 192-201). Heidelberg: Springer.
[3] Grebla, H. A., Cenan, C. O., Distributed Machine Learning based in a Medical Domain. *International Journal of Computers, Communications & Control (IJCCC)*, 2006, Supplem. Volume (*Proceedings of ICCC 2006*, 200-205.
[4] Grebla, H. A., Chiş, G., Stanca, L., Ciumaş, C. (2007). E-learning Platform for Student Recuitment and Retention. In: *Proceedings of IEEII*. A conference on the State of Informatics Education in Europe II, 29-30 November 2007, Thessaloniki, Greece, (pp. 342-351).
[5] Mitrut, M. (2005). *Analiza rețelelor sociale*.Timişoara. Retrieved from http://www.banatbusiness.ro/_files/edit_texte/fisiere/Analiza_retelelor_sociale.pdf.
[6] Neapolitan, R.E. (1990). *Probabilistic Reasoning in Expert Systems*. Wiley & Sons.
[7] Nonaka, I., Takeuchi, H. (1995). *The Knowledge Creating Company*, New York:
[8] Ru-qian, L. (2003). *Knowledge science and computation science*. Beijing: Tsinghua University Press.
[9] Tetsuo, S., Jun, U. et al.(1996). Fusing Multiple Data and Knowledge Sources for Signal Understanding by Genetic Algorithm. *IEEE Transactions on Industrial Electronics*, 43, 411-421.
[10] Tsung-Ting, K., Shian-Shyong, T., Yao-Tsung, L. (2003). *Ontology-Based Knowledge Fusion Framework Using Graph Partitioning*, 2718(2003), 11-20, Berlin / Heidelberg: Springer.
[11] Purdea, I., Pop, I. (2003). *Algebră*. Zalău, Romania Editura Gil.

[12] Purdea, I., Pic, Gh. (1977). *Tratat de algebră, vol. I*. Bucharest: Editura Academică.
[13] Rignet, I. (1948). Relations binaires, fermetures, correspondances de Galois. *Bull.Soc.Math.*, France, 76 (1-4), 114-155.
[14] Sage, A. P., Rouse, W. B. (1999). *Information Systems Frontiers in Knowledge Management*. Information Systems Frontiers 1(3), 205-219.
[15] Smirnov, Al., Pashkin, M., Chilov, N., Levashova, T. & Krizhanovsky, A. (2004). Fusion-Based Intelligent Support For Logistics Management. In: *BASYS, volume 159 of IFIP International Federation for Information Processing* (pp. 209-216), Springer.
[16] Vasilache, D. (2008). *Guvernarea Electronică - O introducere*. Cluj-Napoca: Ed. Casa Cărții de Știință.
[17] Campus area network definition in Wikipedia, from http://ro.wikipedia.org/wiki/Campus_area_network.
[18] State of the Internet, Akaman, from http://www.akamai.com/stateoftheinternet/.
[19] Alun, P. et al. (2000). Kraft: An Agent Architecture for Knowledge Fusion. *International Journal of Cooperative Information Systems*.
[20] Klein, L. A. (2004). *Sensor and data fusion: A tool for information assessment and decision making*. SPIE Press, 2004. p. 51.
[21] Laskey, K.B. (2006). MEBN: A Logic for Open-World Probabilistic Reasoning (GMU C4I Center Technical Report C4I-06-01).
[22] da Costa, P. C. G. & Laskey, K. B. (2005). *Multi-Entity Bayesian Networks Without Multi-Tears* [Draft], DSEOR, Fairfax, USA: George Mason University.