A knowledge management system using Bayesian networks

Article

Accepted version

P. Ribino, A. Oliveri, G. Lo Re, S. Gaglio

In In AI*IA 2009: Emergent Perspectives in Artificial Intelligence, 2009, pp. 446-455

It is advisable to refer to the publisher's version if you intend to cite from the work.

Publisher: Springer

http://link.springer.com/chapter/10.1007%2F978-3-642-10291-2_45

A Knowledge Management System using Bayesian Networks

Patrizia Ribino, Antonio Oliveri, Giuseppe Lo Re, and Salvatore Gaglio

Dipartimento di Ingegneria Informatica (DINFO),
Universita' degli Studi di Palermo, Palermo, Italy
{ribino,antonio.oliveri,lore,gaglio}@unipa.it

Abstract. In today's world decision support and knowledge management processes are strategic and interdependent activities in many organizations. The companies' interest on a correct knowledge management is grown, more than interest on the mere knowledge itself. This paper proposes a Knowledge Management System based on Bayesian networks. The system has been tested collecting and using data coming from projects and processes typical of ICT companies, and provides a Document Management System and an Decision Support system to share documents and to plan how to best use firms' knowledge.

Key words: Knowledge Management, Ontology, Bayesian Network, Decision Support System, Document Management System

DRAFT

1 Introduction

In today's world decision support and knowledge management processes are strategic and interdependent activities in many organizations. Due to the complexity of economy, the firms take measures to change own strategies, structures, technologies and operational mechanisms. Because only who is able to efficiently plan own projects, he will be able to be market competitive. Moreover the great amounts of documents generated during working activities imposes a correct management to avoid loss of important knowledge. In these surroundings, the integration between human processes and computer programs becomes more imperative. The adoption of new computer-based information systems, enabling the storage of structured data and the automation of the information-processing activities of the organization is then fundamental. During the last two decades ad-hoc frameworks known as *Knowledge Management Systems* (KMS) have been proposed, enabling access, coordination and processing of knowledge assets [4][2][3]. KMSs are generally known as information systems which manage organizational knowledge increasing the productivity of operators.

This paper proposes an ontology-based knowledge management framework using two Bayesian Networks in order to support decision maker for project planning and to document management and content analysis. To reach that goal, the system is composed by different modules, an expert system for decision support and a document management engine. The paper is structured as follows: section

2 provides a vision of the state of the art of KMSs and expert systems; in section 3 we introduce essential aspects of KMSs, Bayesian and Decision Networks; in section 4 we present a detailed description of the system; section 5 illustrates the Document Management Service of the KMS. Finally, in section 6, we trace some conclusions.

2 State of the Art

During last decade researches on Knowledge Management and Decision Support has examined several issues about new strategies, tools and systems in order to organize, store and share data and individuals' expertise; at the same time the most important ICT companies have demonstrated an increasing interest in the development of internal knowledge management instruments, using novel data representation models and involving modern AI techniques. A web and ontology-based KMS called WAICENT (World Agriculture Information Centre) is in use on the United Nations Food and Agriculture Organization, in order to improve food security through information use [13]. Liping Sui [12] and Maya Daneva [11] studied the benefits of a decision support system within the business management. In the matter of the decisional systems, N. Fenton et.al [17] and Noothong et.al [18] studied the use of Bayesian networks in order to make decision about software projects. Some architectures for Document Analysis and Understanding was proposed, many of them using Prolog sets of rules like in [15] and in [16]. In this article we introduce a prototypical KMS applied to a real case. By means of the representation of domain concepts through ontologies, two knowledge management mechanisms have been implemented, one related to efficient document management and the other concerning decisional problems about efficient project planning.

3 Overview

3.1 Knowledge Management Systems

Knowledge Management (KM) consists of a technique that uses Information Technology tools for the management of information, and its goal is to improve the efficiency of work teams; it studies methods for making knowledge explicit, and sharing a firm's professional expertises and informative resources. A generic KMS, supporting the creation and storage of knowledge, creates the opportunity to make data, information and knowledge from different sources readily available, managing both explicit and tacit knowledge [5]. To realize such goals, a KMS can make use of different technologies such as: *Document based*, *Ontology/Taxonomy based* and *AI based.*

3.2 Ontologies

An ontology[23] tries to formulate an exhaustive and rigorous conceptual scheme of a particular application domain. Generally it is represented through a hierar-

DRAFT

chical structure which contains all the noteworthy entities, the existing relationships between them, the rules, the axioms and the specific domain constraints. Given a domain of interest, the ontology explains the knowledge structure creating a syntax of domain terms, and shares it with all the people interacting with the given domain.

3.3 Bayesian Network

A Bayesian Network (BN) [21], also called belief network, is a graphical model that shows probabilistic relationships among variables of given problem conditional on uncertainty constraints. In recent years, BNs have been successfully used in many fields such as DataMining [19] and Decision Support System [20]. A BN is a directed acyclic graph with following proprieties:

- A set of random variables are network nodes.
- A set of oriented arcs connect couple of nodes and represent the cause/effect relationships.
- There is a probability table for each node, specifying how the probability of each state of the variable depends on the states of its parents.
- Each node without parents have a prior probabilities table of each state.

The relationship between random variables follows Bayes' rule:

$$p(y|x) = (p(y|x) * p(x)) \diagup p(y) \tag{1}$$

3.4 Decision Network

The Bayesian Networks provide also decision support for a wide range of problems involving uncertainty and probabilistic reasoning. The Decision Network is an extension of BN by adding a decision nodes and a utility nodes. The utility nodes are the variables to optimize and the decision nodes are the variables for decision. The decision node defines a finite set of alternatives corresponding to choices to take in order to achieve the desired aim. The utility values defined in a utility node represent level of preference associated with possible choices. Let C=$c_1, ..., c_n$ be a set of mutually exclusive choices and N be the associated random variables. The objective is reached optimizing the expected utility function that esteems the preferences among the states of the world.

$$EU(c) = \sum_N U(c, N)P(N|c) \tag{2}$$

U(c,N) is the utility value for each configuration of choices and associated random variable. *P(N|c)* represents the probability of N conditioned by choice c, that is when the choice c occurs.

DRAFT

4 Knowledge management for decision support

In this paper we present a KMS prototype for the automation of government office processes, starting from the result obtained using a first release of the prototype called *Kromos* in a real case, that gave the opportunity to measure the increase in information sharing and reuse [1]. This prototype is an ontology-based system of knowledge management with the aim of optimizing business processes for creating and managing ICT projects for generic offices and to obtain a efficient document content analysis and retrieval. To achieve this, the system implements an expert system for decision support and a document management engine. The knowledge representation is based on two ontological domain models, the former reproducing the government offices' structure and the latter modeling the concepts of projects developed by an ICT company. Differently from KMS reported in section 3, our system provides reasoning system and a document management based on Bayesian networks. Figure 1 shows the system architecture.

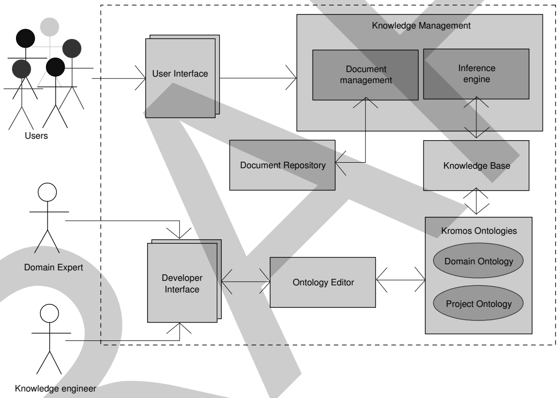


Fig. 1. System architecture

4.1 Knowledge Base and the system Ontologies

The Knowledge Base represents the knowledge container. KB relations and concepts are described using an ontological structure of instances in order to collect and manage data. Ontologies of the proposed system built using Protégé, a free and open source platform developed by Stanford University [7] [8]. Modeling knowledge about the government offices world required some assumptions about its structure and activities, as well as about the nature of the "observer" expected to use and rely on the model. Our ontology can be considered a collection of two correlated ontologies, a domain ontology and a projects ontology; in order to keep the ontology easy to understand, only a few concepts from offices'

DRAFT

domain and from computer engineering projects are collected. This results in a simplified description of Projects, Processes and Structure of offices and a group of attributes and relations. The *Domain Ontology*, representation of the offices' structure and activities, is used to characterize the environment in which the system works. The *Project Ontology* is useful to describe ICT company projects; it maps the structure of the project components containing semi-structured explicit knowledge.

4.2 A Decision network for project planning

The projects planning process is a complicated trial because it is influenced by many different factors. The greatest part of such factors (like as time, costs, resources), being not deterministic, represent sources of uncertainty that must be opportunely esteemed in order to optimize the decisional trials of business planning. In the last years, the BN has become a popular representation for encoding uncertain expert knowledge in expert system [22]. In this section, a model of decision support system based on BNs is proposed. The development of the net has been divided in three phases: domain analysis, relationships discovery among the variables of interest and estimate of the probability table. The obtained decision network is shown in fig.2.

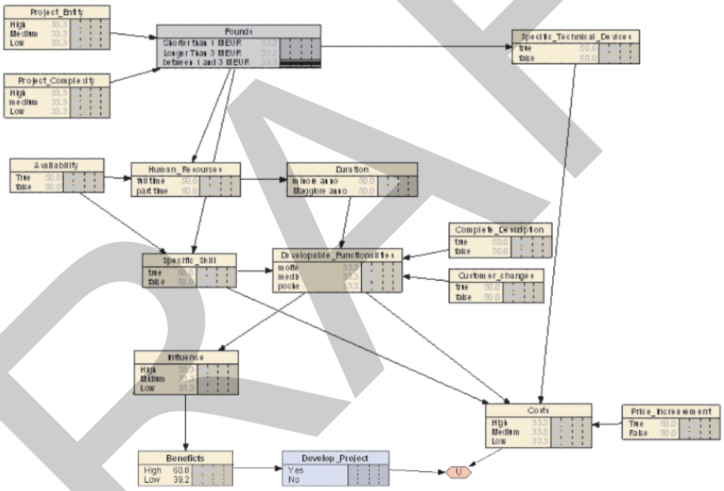


Fig. 2. Projects decision network.

Domain Analysis It concerns the determination of all variables that characterize the domain of interest and the individuation of all the possible states associated of each variable of the network. Each random variable represent a node of the BN. The set of variables discovered are the followings:

DRAFT

- *Project Entity (PE)*: the number of modules whose it is constitute. States Values: High, Medium and Low.
- *Project Complexity(PC)*= how the project is intricate. States Values: High, Medium and Low.
- *Founds*(F)= available financing in Million of Euro (MEUR): States Values: Shorter than 1 MEUR, Between 1 and 3 MEUR and More than 3 MEUR.
- *Specific Technical Devices (STD)* = availability of specific technical devices: States Values: True and False.
- *Availability(A)* = availability of human resources. States Values: True and False.
- *Human Resources (HR)* = availability of employees. States Values: Full Time and Part Time.
- *Duration (D)* = project duration: States Values: High and Low.
- *Specific Skills (SS)* = availability of human resources with specific technical Know How. States Values: True and False .
- *Develop Functionalities (DF)* = project functionality that can be implemented. States Values: Many, Mean and Few.
- *Complete Description (CD)* = if the customer has provided an exhaustive description of the product. States Values: True and False.
- *Customer Changes (CC)* = if the customer can bring changes to the requisite of the product in work progress. States Values: True and False.
- *Influence (I)* = how much the project is important. States Values: Many, Mean and Few.
- *Costs (C)* = total costs for the development of the project. States Values: High, Medium and Low.
- *Price Increment (PI)* = Prices growing (raw material, renewal employment contract etc.). States Values: True and False.
- *Benefits (B)* = project benefits (improvement of human process, of time and costs). States Values: High and Low.

Relationships discovery The Relationships discovery phase gave the chance to discover the causal relationships between the variables objects of our observation, how they influence, or how they are influenced by, other variables. Examining the domain to be modeled, this set of dependence and independence conditions were discovered:

$$\begin{aligned} PE, PC, PI, CD, CC, A|\varnothing; \quad F|PE, PC; \quad HR|A; \\ SS|A, F; \quad D|HR; \quad DF|SS, CD, CC; \\ I|DF; \quad C|SS, DF, STD, I; \quad B|I; \end{aligned} \tag{3}$$

For instance, in this equations the symbol | represents the dependence of the right-side set of variables by the left-side set of variables.

DRAFT

Estimate of the probability tables To define the tables of conditional probability distributions for each node of the BN is generally the hardest task. In our specific case, this difficult is tied to the experience of the domain expert because every specialist could have been managed only some projects cases that could be a non-realistic sample set for this calculus. To avoid that, we can consider not only the opinion from one expert of the domain, but the different opinion about the same node coming from different domain experts, to reach a unique final opinion expressed as probability. To intersect all this experience data in a unique value of probability, the schema in figure was adopted:

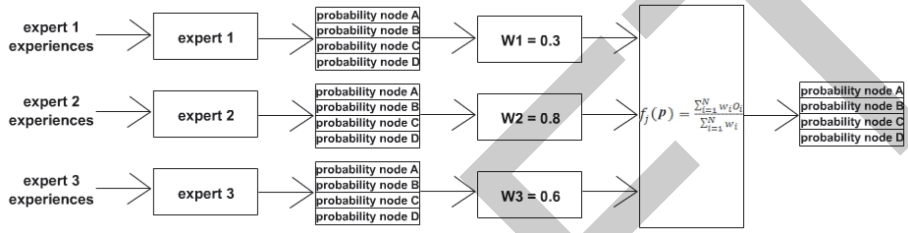


Fig. 3. Fusion schema of experts opinion.

The opinions about node j, produced by N experts can be combined using a weighed average:

$$f_j(o) = (\sum_{i=1}^{N} w_i o_i)/(\sum_{i=1}^{N} w_i) \tag{4}$$

In this equation O_i indicates the percentage value of the opinion of the expert i, while w_i means the related weight, calculated as it follows:

$$w_i = G_i * A_i/(G_{max} * A_{max}) \quad with \quad G \in [1, G_{max}] \quad and \quad A \in [1, A_{max}] \tag{5}$$

where G indicates the degree of experience of the expert in the company, while A indicates the number of years he worked for that comany.

5 Document Management Service

In ICT companies the volume of documents produced during working activities grows rapidly. Documents contain most of the information about projects, functionalities, people involved and so on, so it is necessary to develop automatic system that, starting from the analysis of the document structure, could understand the content and organize information in a structured way. The first step for the understanding of a document is the *Document Structure Analysis* performed by a *Specification Module*, that can capture data about the physical and

DRAFT

logical layout of the document described by a set of rules. The second step is the *Document Content Analysis* performed by an *Extraction Module*, so that Project Ontology instances can be enriched capturing information about projects, and Domain Ontology instances can be enriched adding relation instances between project and offices that will use the products realized, and which functionalities will be automatized by the project described in the document.

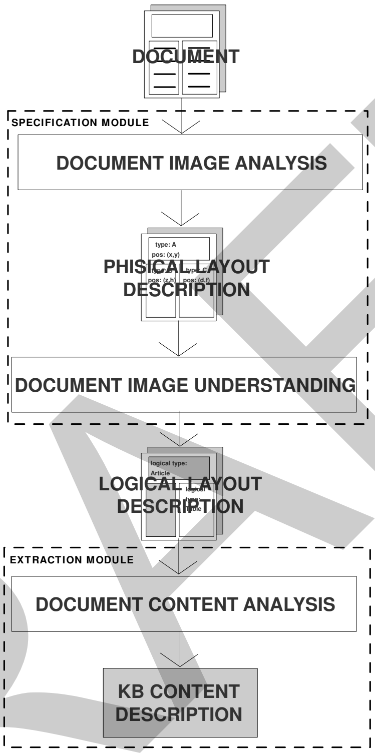


Fig. 4. Document Management Service Architecture.

Our platform performs this Document Management in different steps, as fig. 4 shows.

5.1 Specification Module

Document Geometric and Logical Layouts. A common document can be considered as a set of textual objects (text elements such as titles, subtitles, captions) and graphical elements (images, separation lines, tables); as [15][14] suggest, blocks are the smaller parts of the document structure and form the *Document Geometric Layout* (DGL), that can be recognized using *Document Image Analysis* techniques (such as Optical Character Recognition instruments)

DRAFT

obtaining a digital description of the physical segmentation of the document. The most important phase of the process is the definition and recognition of the *Document Logical Layout* (DLL), performed using *Document Image Understanding* techniques. While Geometric Layout recognition extracts geometric structures of the document, Document Image Understanding maps the geometric structure in a logical structure, considering the logical relationships between basic blocks using information like position, dimension, type of blocks. In order to collect information about geometric and logical layout we used a Structure Ontology of instances, mapping blocks definition in two concepts: *block*, that collects the information about a single block and the logical unit it belongs to, the dimension and position in the document, the previous and subsequent blocks in the unit; and *content*, that represents information about the typology of the block (textual or graphic), and the content itself to be used in the subsequent content analysis. Each instance of block is connected to one instance of content class, so that it is performed a separation between layout and content data. Some of this information become from the Document Analysis, other from the subsequent Document Understanding phase.

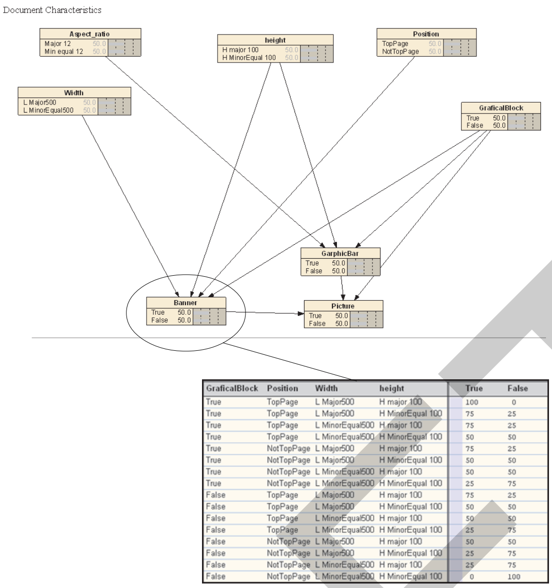
DRAFT

Logical description rules and the Bayesian recognition. In ICT companies greatest part of documents are structured or semi-structured, so that a domain expert can define their logical structures using a set of unary or multiple rules to recognize logical units, based on the analysis of some features considered necessary for the identification of such relationships. In the DMS we present, all the rules are described using Bayesian Networks, used to model the statistical dependency between variables and the local probability distributions of leaf nodes related to father nodes, representing the joint probability distribution of a set of features. Each feature is a node of the net that can assume two or more discrete values, connected to nodes that represent the logical units of the document. Oriented arcs connect feature nodes and logical-unit nodes, the latter provided with a table of conditional probability distributions, mapping the probability that, given a set of feature values, the blocks can be recognized as a part of a defined logical unit. Probability values of a node can be determined as it follows:

$$Pr = 100 * (N' \diagup N) \tag{6}$$

considering N, the number of features from witch the category-node depends and N' the number of requirements satisfied by the block. For instance, fig. 5 shows the part of the net regarding the graphical components of a document.

A graphical geometric block can be considered, as logical units, a banner, a graphical bar or a picture unit, while the features we have to consider are that it is a geometric graphic block, its hight and width and its position. In the same way all the other logical components of the document we want to analyze are described using the BN. The entire set of conditional probabilities of the net guarantees that, having information about the characteristics of all the geometric blocks of the document, the net can identify the logical units they



GraficalBlock	Position	Width	height	True	False
True	TopPage	L Major500	H major 100	100	0
True	TopPage	L Major500	H MinorEqual 100	75	25
True	TopPage	L MinorEqual500	H major 100	75	25
True	TopPage	L MinorEqual500	H MinorEqual 100	50	50
True	NotTopPage	L Major500	H major 100	75	25
True	NotTopPage	L Major500	H MinorEqual 100	50	50
True	NotTopPage	L MinorEqual500	H major 100	50	50
True	NotTopPage	L MinorEqual500	H MinorEqual 100	25	75
False	TopPage	L Major500	H major 100	75	25
False	TopPage	L Major500	H MinorEqual 100	50	50
False	TopPage	L MinorEqual500	H major 100	50	50
False	TopPage	L MinorEqual500	H MinorEqual 100	25	75
False	NotTopPage	L Major500	H major 100	50	50
False	NotTopPage	L Major500	H MinorEqual 100	25	75
False	NotTopPage	L MinorEqual500	H major 100	25	75
False	NotTopPage	L MinorEqual500	H MinorEqual 100	0	100

Fig. 5. Particular of the Bayesian Network used for Logical Layout recognition.

represent. This information complete the instances in the KB, so that textual analysis can be accomplished.

5.2 Extraction Module

The extraction module we implemented uses Apache Lucene, an Information Retrieval (IR) engine adapted for the analysis of logical units contents. Lucene is an open-source, high performance, scalable, full-featured text search engine and information retrieval library written in Java and suitable for any application requiring a full text search [6], through which any piece of data convertible to a textual format can be indexed and made searchable [9]. Searches in Lucene are performed specifying one or more keywords and one or more fields to search within. Search results (in the form of Lucene Documents) are collected within Lucene Hits. Since the Extraction Module is based on Lucene, it possesses similar advantages: very fast response time and almost hidden complexity to users. In the actual release of the module, it is performed: a content elicitation to withdraw textual content eliminating irrelevant information, such as typesetting format, and transforming it into a character data stream; and a content tokenization that breaks the content into words and sentences according to lexical analysis, transforming the data stream into a set of terms for the subsequent content parsing procedure. After that, Apache Lucene performs an indexing activity to the logical unit contents; in Apache Lucene an index can be considered as a sequence of documents, which are a sequence of fields [10]. This activity is absolutely transparent to the user. After the initial collection of information

DRAFT

by the documents and the creation of an index, users need to retrieve this data, seeking them using a search interface. The Lucene system searches in the indexed data and retrieves the relevant information, using a set of different factors such as term frequency and inverse document frequency, considering only those units that it believes are relevant.

6 Conclusion

In the new economy knowledge and its efficient exploitation has become a key factor for organizational success, pressing the organizations to adapt themselves to changing environment. In this regard, Knowledge Management strategies are promoted and several information systems are developed to give support to knowledge processes. The key thrust of this article has been to analyze the benefits of Bayesian Networks and Ontology based systems for knowledge management. The system has the aim to improve the growth of organizational knowledge for projects' management, developing a KMS prototype which makes use of an Expert System for decision support and a Document Management System for document analysis. The opportunity to adapt this KMS architecture modeling different domains, gives the chance to use it to different application contexts and increasing information sharing and reuse.

DRAFT

References

1. Oliveri, A. and Ribino, P. and Gaglio, S. and Lo Re, G. and Portuesi, T. and La Corte, A. and Trapani, F.: KROMOS: ontology based information management for ICT societies, ICSOFT (2009)
2. Staab, S. and Studer, R. and Schnurr, H.P. and Sure, Y.: Knowledge Processes and Ontologies, IEEE Intelligent Systems, 26–34 (2001)
3. O'Leary, D.E.: Enterprise Knowledge Management, Computer, IEEE Computer Society, 54–61 (1998)
4. Alavi, M. and Leidner, D.E.: Review: Knowledge Management and Knowledge Management Systems: Conceptual Foundations and Research Issues, Knowledge Management, Routledge (2005)
5. Takeuchi, H. and Nonaka, I.: The knowledge-creating company: How Japanese companies create the dynamics of innovation, Oxford University Press, NY (1995)
6. Gospodnetic, O. and Hatcher, E.: Lucene in Action, Manning (2005)
7. Minsky, M.: A Framework for Representing Knowledge, Massachusetts Institute of Technology Cambridge, USA (1974)
8. Chaudhri, V.K. and Farquhar, A. and Fikes, R. and Karp, P.D. and Rice, J.P.: OKBC: a programmatic foundation for knowledge base interoperability, Proc. of the 15 national conf. on AI/Innovative applications of AI, 600–607 (198)
9. Pirro, G. and Talia, D.: An approach to Ontology Mapping based on the Lucene search engine library, Proceedings of the 18th Int. Conf. on Database and Expert Systems Applications, USA, Vol. 00, 407–411 (2007)
10. Bennett, M.: Contrasting relational and full-text engines, NIE Enterprise Search Newsletter 2 (2004)

11. Daneva, M. and Peneva, J. and Rashev, R. and Terzieva, R.:Knowledge-Based Decision Support System for Competitive Software Audit, IEEE Int. Conf. on Systems Man and Cybernetics, 3, 1974–1979 (1995)
12. Sui, L.:ecision support systems based on knowledge management, Services Systems and Services Management, Proc. of 2005 Int. Conf. on, 2 (2005)
13. O'Leary, D.E.: A multilingual knowledge management system: A case study of FAO and WAICENT, Decision Support Systems, Elsevier, 45, 641–661 (2008)
14. Klink, S. and Dengel, A. and Kieninger, T.: Document structure analysis based on layout and textual features, Proc. of Int. Workshop on Document Analysis Systems, DAS2000, 99–111 (2000)
15. Niyogi, D. and Srihari, S.N.: Using domain knowledge to derive the logical structure of documents, Proc. The International Society for Optical Engineering, 114–125 (1996)
16. Dengel, A. and Barth, G.: ANASTASIL: A hybrid knowledge-based system for document layout analysis, Proc. of 11th IJCAI, The William Kaufmann Inc, 1249–1254 (1989)
17. Fenton, N. and Marsh, W. and Neil, M. and Cates, P. and Forey, S. and Tailor, M: Making resource decisions for software projects, Software Engineering, ICSE 2004. Proc. 26th Int. Conf. on, 397-406 (2004)
18. Noothong, T. and Sutivong, D.: Software Project Management Using Decision Networks, Intelligent Systems Design and Applications, 2006. ISDA'06. 16 Int. Conf. on, Volume n.2, (2006)
19. Heckerman, D.: Bayesian networks for data mining, Data mining and knowledge discovery,79–119 (1997)
20. Zhang, SZ and Yang, NH and Wang, XK: Construction and application of bayesian networks in flood decision support system, Proc. of the First Int. Conf. on Machine and Cybernetics, 718–722 (2002)
21. Heckerman, D. and others: A tutorial on learning with Bayesian networks, NATO ASI SERIES D BEHAVIOURAL AND SOCIAL SCIENCES, 301–354 (1998)
22. Heckerman, D. and Mamdani, A. and Wellman, M.P: Real-world applications of Bayesian networks, Communications of the ACM, 24–26 (1995)
23. Uschold, M. and Gruninger, M.:Ontologies: Principles, methods and applications, Knowledge engineering review, 1996

DRAFT