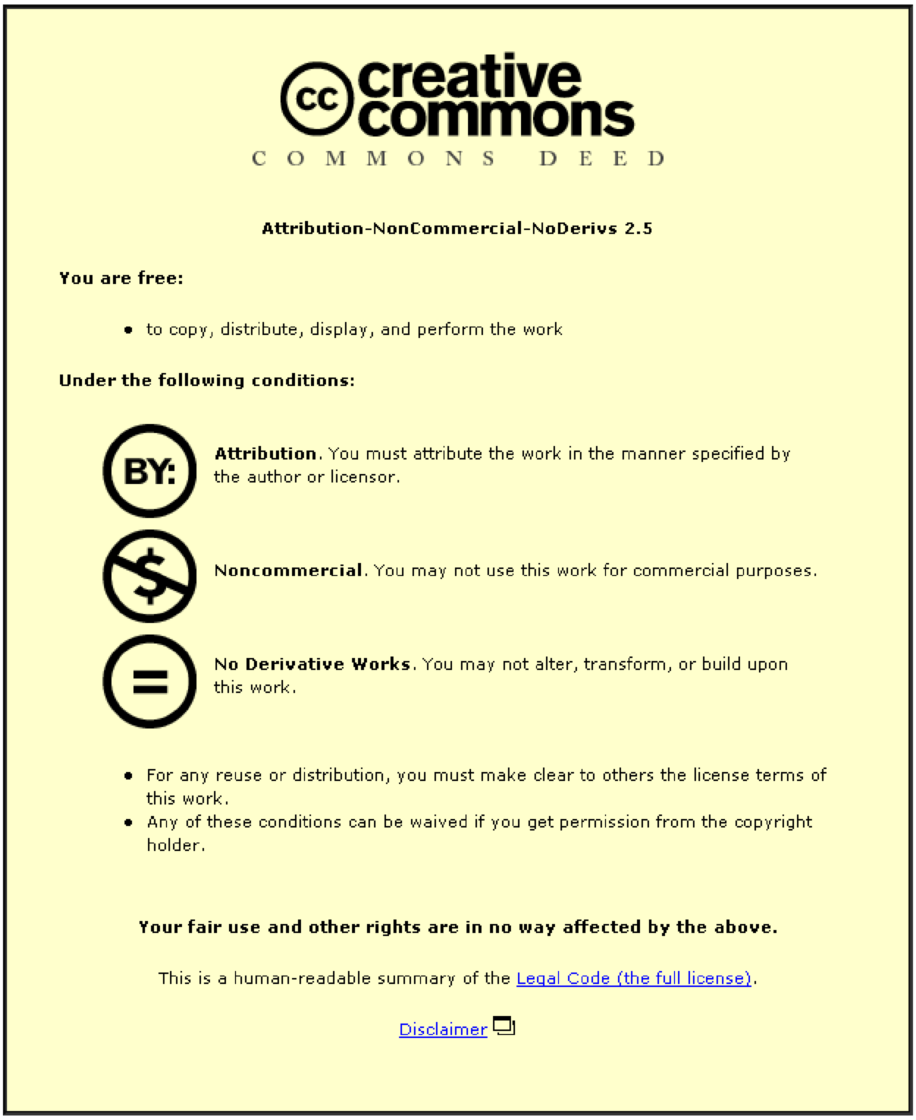


Generating Rules from Data mining for Collaboration Moderator Services

C. Palmer[1], J.A. Harding[1]*, R. Swarnkar[1], B.P. Das[1], and R.I.M. Young[1]

[1]: Wolfson School of Mechanical and Manufacturing Engineering
Loughborough University, Leicestershire, U.K., LE11 3TU

*Corresponding Author: J.A.Harding@lboro.ac.uk

Tel: +44(0)1509227662

Fax:+44(0)1509227648

Email: J.A.Harding@lboro.ac.uk

ABSTRACT

A Moderator is a knowledge based system that supports collaborative working by raising awareness of the priorities and requirements of other team members. However, the amount of advice a Moderator can provide is limited by the knowledge it contains on team members. The use of data mining techniques can contribute towards automating the process of knowledge acquisition for a Moderator and enable hidden data patterns and relationships to be discovered to facilitate the moderation process. A novel approach is presented, consisting of a knowledge discovery framework which provides a semi-automatic methodology to generate rules by inserting relationships discovered as a result of data mining into a generic template. To demonstrate the knowledge discovery framework methodology an application case is described. The application case acquires knowledge for a Moderator to make project partners aware of how to best formulate a proposal for a European research project by data mining summaries of successful past projects. Findings from the application case are presented.

Keywords: Collaborative projects, Moderators, Knowledge discovery, data mining, virtual enterprise, virtual organization

1. Introduction

The collaboration moderator service (CMS) (Swarnkar et al, 2011) is one of the collaborative knowledge services developed by the SYNERGY project, which is funded by the European Commission under Grant ICT-2007.1.3-216089. The SYNERGY project aims to provide an infrastructure and services to networked SMEs (small to medium sized enterprises) to discover, capture, deliver and apply knowledge relevant to forming and participating in virtual organizations (VOs). A VO has been defined as: “A temporary alliance of independent enterprises that come together to share skills, core competencies and resources in order to better respond to business opportunities and whose cooperation is supported by computer networks” (Luczak and Hauser, 2005).

Every VO is unique and the set of activities a VO undertakes to achieve its business objectives are also unique, as demonstrated by Mezga´r and Kova´cs (1998). However, VOs can benefit from the experiences of previous projects with similar team members or objectives. VO members need to be made aware of the potential problems that can occur during VO operation in a timely manner without the project manager suffering from information overload (Nof, 2006). A

Moderator is a knowledge based system that supports collaborative working by raising awareness of the priorities and requirements of other team members (Harding et al, 2003). The Synergy project is developing a moderator service (CMS) to support VOs.

The amount of advice a Moderator can provide is limited by the knowledge it contains on team members and their roles. A moderator requires knowledge about each member's area of interest, competencies, current activities and the actions that need to be taken when changes that affect these activities occur. Currently all knowledge acquisition for prototype moderator software is performed manually, based on human expertise and experience. To conform to good business practice enterprises document their operations and business outcomes and this documentation has the capability to provide a knowledge resource for a Moderator. Business cases, progress reports, project logs, change logs, quality reports and post project reviews can all be useful sources of knowledge for learning about good (or poor) practices or experiences. Knowledge can be generated from an enterprise's documentation by applying data mining approaches to extract patterns, relationships and discover rules and this type of learning can then potentially be applied to advise the activities of future projects. In addition to contributing towards the automation of the knowledge acquisition process for a Moderator, the use of data mining techniques enables hidden data patterns and relationships to be discovered to facilitate the moderation process. These relationships may have remained undetected due to their complexity or because of the huge amount of information available.

The SYNERGY project aims to provide services for VOs throughout their lifecycle, and the lifecycle of a VO has been defined in the SYNERGY project through four distinct phases: Pre-creation, Creation, Operation and Termination. In the Pre-creation stage potentially collaborative enterprises need to be made aware of possible business opportunities and of other potential collaborators. In the Creation stage, stronger partnerships are made between the companies through the formation or creation of a virtual organization in response to a new business opportunity. During the Operation stage the project deliverables are achieved for which the VO was created. Upon Termination of the VO, any useful knowledge and learning from the VO needs to be maintained for the future use of the participating individual enterprises and future VOs. Knowledge and learning should therefore be transferred to the individual members as central knowledge stores are dissolved. Moderation activities and knowledge acquisition for a Moderator can occur at any stage during the lifecycle.

This paper considers the possibility of using the knowledge acquired from data mining techniques to semi-automatically generate rules for a Moderator. These rules will form part of the Moderator's knowledge base and will be used to make team members aware of potential problems occurring during the operation of a project. The rules are generated by inserting knowledge patterns discovered from data mining into a generic rule template. The novel approach presented in this paper combines data mining techniques with rule generation, reducing the time required for conducting knowledge acquisition interviews. However, the rules generated need to be verified by a domain expert to ensure that they are applicable. In addition, the expert may also chose to derive other rules from the knowledge patterns and relationships discovered. The combination of automatic rule generation with expert verification and expert derived rules results in a semi-automatic knowledge acquisition process, decreasing the amount of expert time needed for knowledge maintenance and update.

The paper commences by providing some contextual information on data mining. An overview of collaborator moderator services is also given. To implement data mining within CMS a knowledge discovery framework is presented. The proposed framework will provide a semi-automatic methodology to generate rules for CMS. A generic template which enables rules to be discovered is presented. To demonstrate the knowledge discovery framework methodology an application case is described. The application case focuses on the VO Pre-creation stage, acquiring knowledge for CMS to make project partners aware of how to best formulate a proposal for a European research project by data mining summaries of successful past projects. Results achieved by applying the knowledge discovery framework to the application case are given. Discoveries found by applying the methodology are described and possibilities for future work are discussed.

2. Knowledge Discovery and Data Mining

Data mining is one of the steps in the process of discovering knowledge from data (KDD). Other process steps include data preparation, data selection, data cleaning (to remove noise and inconsistent data), incorporation of appropriate prior knowledge and interpretation of the data mining results (Fayyad et al, 1996). Data mining comprises the application of intelligent algorithms to extract patterns from data (Fayyad et al, 1996; Han and Kamber, 2006) and covers a wide range of different tools and techniques ranging from statistics to Artificial Intelligence. This research makes generic use of the terms "knowledge discovery" and "data mining" to illustrate how a form of semi-automatic knowledge discovery can be used in CMS. A brief overview of data mining is presented within this context. More detailed reviews of data mining research are given by Choudhary et al (2009), Wang (2007) and Harding et al (2006).

Choudhary et al (2009a) consider the use of text mining applications to extract knowledge from post-project reviews. Text mining applies a subset of data mining algorithms to uncover patterns across sets of unstructured textual data. The information contained in post project reviews enables organizations to learn from previous projects. Some of the techniques demonstrated by Choudhary et al (2009a) are expanded on to derive patterns and relationships for the research described within this paper.

Grobelnik and Mladenic (2003) use text mining methods to analyze European projects, which is the same application area considered by this paper. Text mining methods were used to group the projects according to their content and the organizations participating in the projects. The research concentrates on the data mining step of the KDD process with only a limited amount of data preparation. This restricted the amount of information which could be discovered from the data, as without data preparation only a limited number of concepts are available to analyze. The research identifies similar organizations and project topics; analyzes collaborating organizations and inter-organization connectivity; and identifies consortia of organizations for a given topic. The knowledge discovered is not utilized although the suggestion is made that in the future potential new organizations for proposed research projects could be identified.

The work of Grobelnik and Mladenic (2003) was supported by the SolEuNet European research project (IST-1999-11495) which enabled a group of academic and business teams to form a virtual organization. SolEuNet considered data mining, decision support and integrating these two technologies to solve new types of problems. Decision support "is concerned with developing systems aimed at helping decision makers solve problems and make decisions"(Mladenic et al, 2003). Other relevant work within the project discusses the use of data mining to support decision making. Lavrac and Bohenec (2003) describe using data mining to provide information for a housing loan application. Gasar et al (2003) develop models for the prediction of high school grades by the sequential application of data mining and decision support methods. In both applications the decision models are developed manually.

Other work considering the integration of data mining and decision support relevant to this paper is the Intelligent Decision Support System (IDSS) described by Heath (2006) and the knowledge-based approach proposed by Lima et al (2010). Heath extends the CRISP-DM knowledge discovery process model (Shearer, 2000) for use with medical datasets. The extended model assists in the automated extraction of domain knowledge from existing patient data and could be used to populate the knowledge component of a proposed decision support system.

Lima et al (2010) describe a knowledge-based model which provides functionalities to share best practices and lessons learned from previous projects. The model is validated under the scope of the CoSpaces project (IST-5-034245). CoSpaces is a European research project which considers a software framework to support collaboration within distributed manufacturing enterprises. The approach intends to support teams by enhancing decision making in co-located and distributed project meetings, through the anticipation of problems, deviations and solutions. The aim is to aid project teams by providing them with relevant historical cases from previous projects and information of how these cases were solved. The system is based on ontology-based classification and indexation of similarities among projects; and data mining of the issues, decisions and solutions in past projects. A mining services component is used to identify patterns of problems and solutions and the relationships between them. The main capabilities proved by the mining services are classifying the data, grouping facts into similar clusters and discovering data patterns.

None of the above applications are able to generate models based on the results of data mining for a project advisory system. All the applications recognize the potential of this technique in acquiring information to advise a user. Heath (2006) describes a proposed system to integrate data mining and decision support. Lavrac and Bohenec (2003) and Gasar et al (2003) utilize the results of data mining to manually create decision models. The work of Lima et al (2010) is an ongoing research under validation. No examples are provided to demonstrate how much of the approach has been implemented.

3. Collaboration Moderator Services Structure

The Collaboration Moderator Service (CMS) develops earlier Moderator research (Harding et al, 2003; Lin et al, 2005) and applies it to support VOs. The aim of CMS is to raise awareness of opportunities or potential problems and the needs, possible consequences and likely outcomes in collaboration activities between the partners of a VO. The architectural design of CMS is shown in figure 1 below. The main modules will be briefly described here, for more details see (Swarnkar et al, 2011).

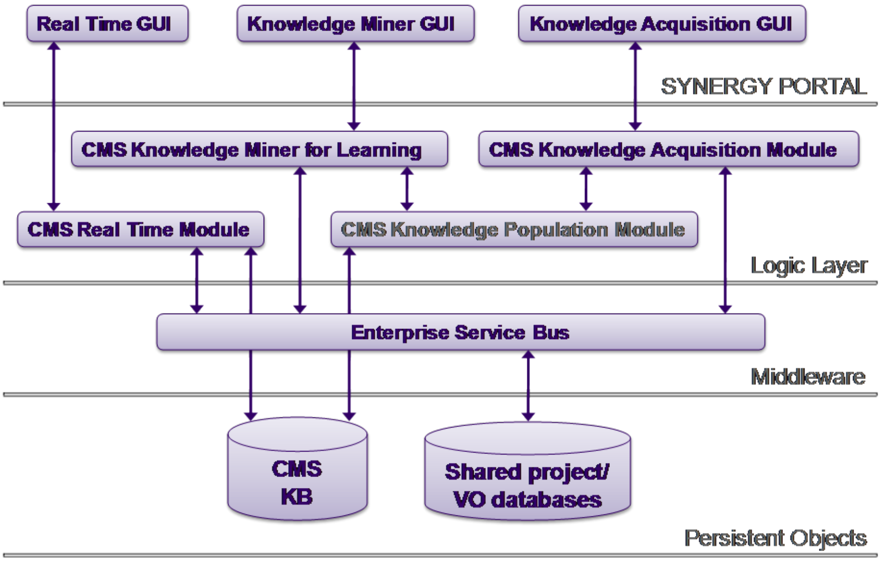

Real Time GUI
Knowledge Miner GUI
Knowledge Acquisition GUI
SYNERGY PORTAL
CMS Knowledge Miner for Learning
CMS Knowledge Acquisition Module
CMS Real Time Module
CMS Knowledge Population Module
Logic Layer
Enterprise Service Bus
Middleware
CMS KB
Shared project/ VO databases
Persistent Objects

Figure 1

The CMS design contains several user interfaces (Real Time GUI, Knowledge Miner GUI, Knowledge Acquisition GUI) to allow a human user to interact with its modules. Below the interfaces lays the logic layer containing CMS Knowledge Miner for Learning, CMS Knowledge Acquisition Module, CMS Real Time Module and CMS Knowledge Population Module.

The CMS Knowledge Miner for Learning performs mining activities to identify new knowledge and update existing knowledge. The CMS Knowledge Acquisition Module obtains and generates the knowledge required by CMS to support its users and uses this knowledge to populate the CMS knowledge base (CMS KB). More details of how the CMS Knowledge Miner for Learning and CMS Knowledge Acquisition Module implement data mining are given in the next section "Knowledge Discovery Framework". The CMS Real Time Module monitors the flow of events within the VO, identifying when one or more of the VO members needs to be made aware of potential opportunities or problems that exist within the VO by comparing the events with VO member information contained in the CMS KB. The CMS Knowledge Population Module provides services to update, add and delete knowledge from the CMS KB. In the prototype implementations of CMS it has been incorporated in CMS Knowledge Acquisition Module.

The Enterprise Service Bus is comprised of an integration platform which connects and coordinates the interaction of the SYNERGY services and enables message passing and data transformation. The Enterprise Service Bus uses messages to signal events and facilitates data sharing and knowledge base access by SYNERGY services. In the SYNERGY project this is provided by Petals (Petalslink.com).

In the persistent objects layer, the CMS KB contains information about the VO members and also knowledge on how to perform the moderation activities. The knowledge for moderation activities is stored within a set of expert modules. The VO shared project information is accessed through one of the other SYNERGY services, providing knowledge on enterprise core competence, process knowledge for VO formation and member selection and VO operations management knowledge. To ensure correct sharing and protection of knowledge, knowledge bases and services within the shared project information are structured using an ontology called ECOS (Khilwani, 2011; Khilwani et al, 2011).   The ECOS ontology classifies enterprise information (competence, management, etc.) in an explicit manner.

4. Knowledge Discovery Framework

To implement data mining within CMS a knowledge discovery framework has been devised based on the KOATING framework of Choudhary et al (in press). The proposed framework will provide a semi-automatic methodology to discover and utilize new knowledge, thus reducing the amount of expert time required for knowledge discovery. The framework consists of knowledge miners, a repository, multiple expert modules and a knowledge manager.

The knowledge miners use a range of knowledge discovery and data mining tools to extract relationships and useful knowledge from the shared VO databases. A knowledge miner can be domain specific, allowing a high degree of customization, or specific to a data analysis technique.   Knowledge miners can be implemented in a variety of ways, e.g. as software code or as an expert system.

The repository provides the knowledge sets required by knowledge miners and temporarily stores the mining results and mining parameters such as confidence levels. Each of the expert modules manually stores knowledge about a team /VO member and semi-automatically stores knowledge discovered as a result of data mining.

The knowledge manager mediates requests for data and manages information flow between the various knowledge miners, the expert modules and the repository. Upon receiving a mining request for information, the knowledge manager first queries the repository to see if knowledge pertaining to the request already exists. If the knowledge is not contained in the repository, the knowledge manager initiates the knowledge miner(s) to mine the appropriate project/VO knowledge/data bases. When mining is complete the knowledge manager updates the repository and deploys the new or updated knowledge discovered into the relevant expert modules.

Figure 2 below shows how the proposed knowledge discovery framework can be integrated with CMS.

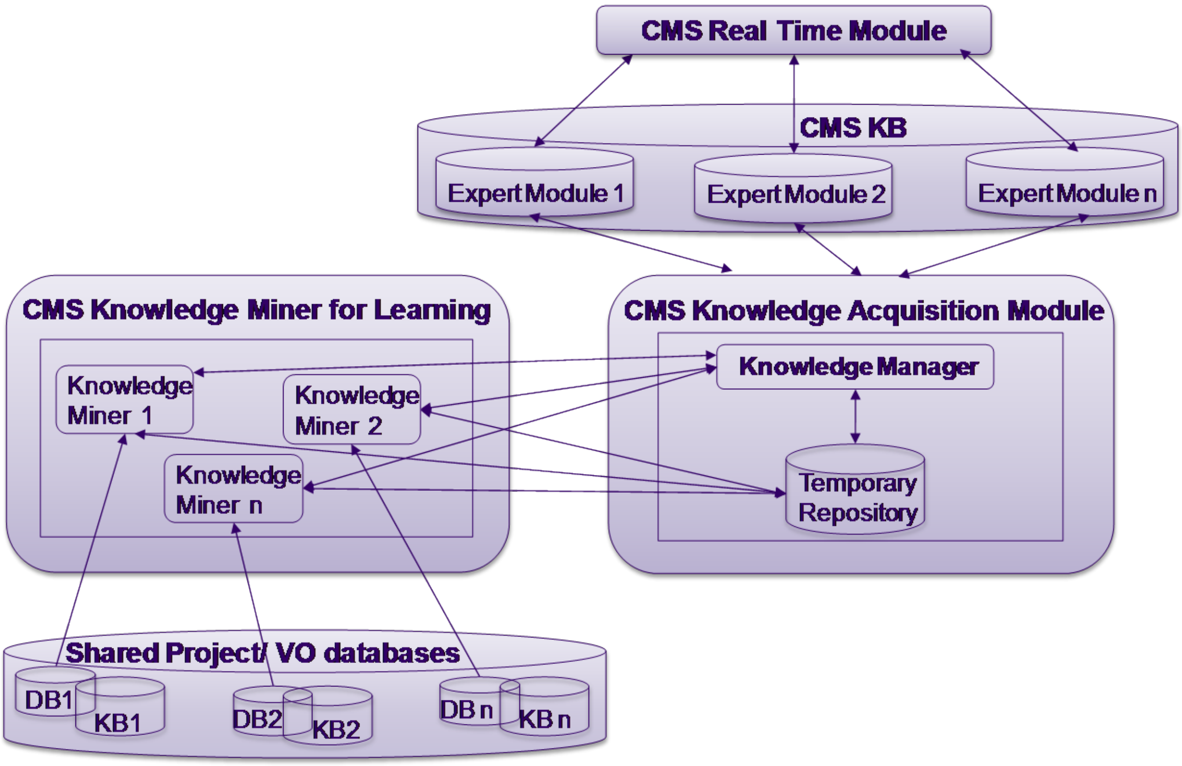


Figure 2

The process model implemented by this research is based on the Knowledge Discovery Module process model of Choudhary et al (in press) which modified the CRISP-DM process model (Shearer, 2000). The knowledge discovery process model consists of seven phases: domain understanding, data understanding, data preparation, application of algorithms and evaluation of results, deployment and conflict resolution. Since a phase may require reassessing as a result of information discovered in a subsequent phase, the phases do not have a defined sequence order.

Domain understanding is needed to interpret data. Understanding data, its structure and meaning, is needed for data preparation. Conversely, understanding of data leads to better domain understanding. Data preparation consists of structuring the data and transforming it into data sets to which algorithms may be applied. The algorithms extract metadata or high-level information and locate patterns and relationships within the extracted information. Further algorithms can be applied to these patterns and relationships to derive rules. The application of algorithms creates more data which may require structuring (returning to the data preparation phase). The evaluation phase assesses whether the generated knowledge is novel and useful within its application domain. Statistical techniques may be employed to determine the importance of the knowledge. Relationships may be evaluated for statistical significance before revisiting the application of algorithms phase to derive rules from these relationships. During deployment the discovered knowledge is put into use. Conflicts may occur between the newly discovered knowledge and existing knowledge. Conflict resolution requires input from a domain expert. Figure 3 depicts the phases of the knowledge discovery process model and shows how they interact with framework elements.

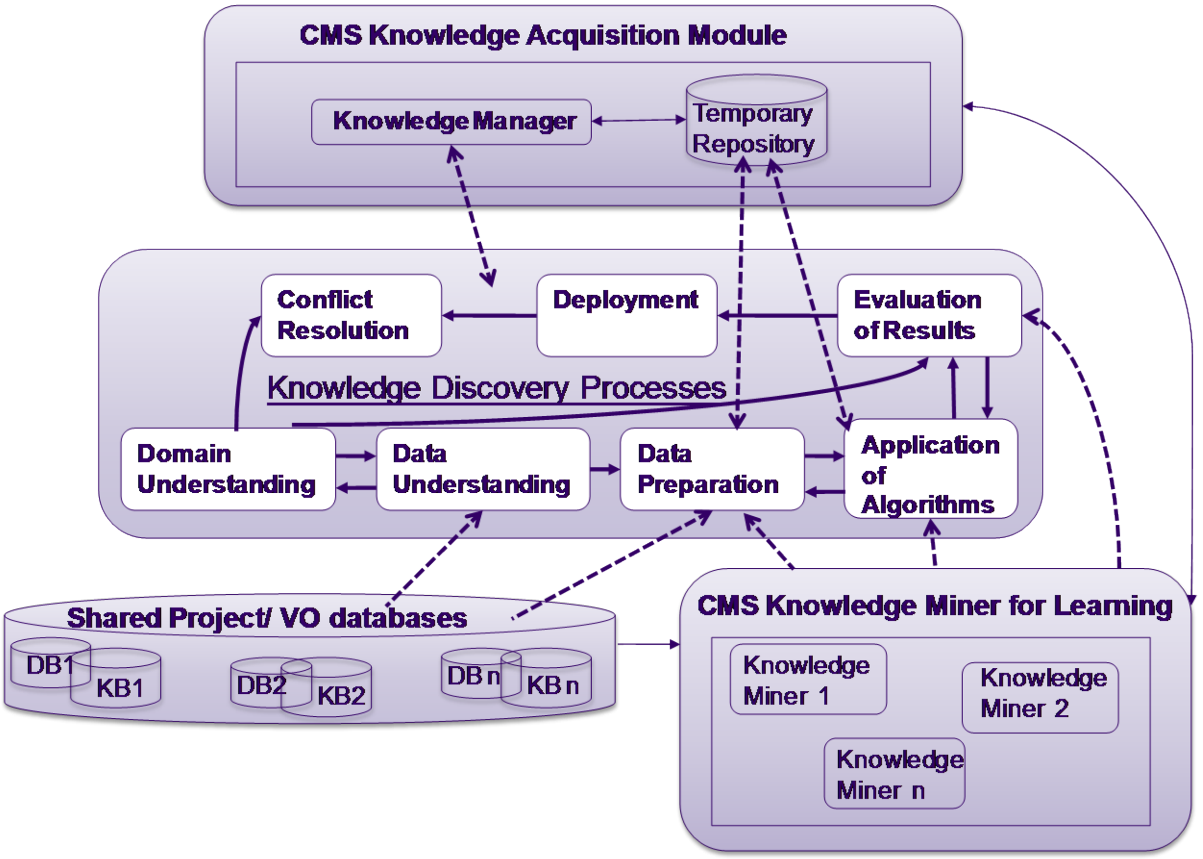


Figure 3

5. An Application Case

To demonstrate the knowledge discovery framework methodology to acquire knowledge from data mining techniques and utilize this knowledge to semi-automatically generate rules for CMS an application case has been developed. The VO business opportunity considered by the application is European research projects, of which the SYNERGY project is an example. The Seventh Framework Programme (FP7) funded European research projects that consist of consortia (usually digitally supported) of universities, research centres and SMEs collaborating to achieve research objectives, and hence these project teams can be classed as VOs. The application case focuses on the VO Pre-creation stage, acquiring knowledge for CMS to make project partners aware of how to best formulate a project proposal by using data mining to detect the “hidden” features of successful past projects. The data sources consisted of one page project summaries of previous projects. The research objective chosen was Challenge 1 “Pervasive and Trustworthy Network and Service Infrastructures”, which resulted in 142 summaries (downloaded from http://cordis.europa.eu/fp7/projects_en.html, activity ICT-2007.1, 16/08/10. In the context of SYNERGY this web source can be regarded as one type of shared VO databases).

To assess the feasibility of the semi-automatic methodology and to provide some initial results the application utilized PolyAnalyst 6.0, a commercial standalone software package. Future development is required to assess suitable open source data mining tools to incorporate within CMS to achieve the functionality of the procedures discovered using PolyAnalyst. This will enable CMS to apply the knowledge acquired through data mining.

This section describes how the knowledge discovery framework is implemented by examination of the application case and what activities are employed within the phases of the knowledge discovery process model to use the knowledge to semi-automatically generate rules.

5.1. Application Framework

As described in section 4 the knowledge discovery framework consists of knowledge miners, a repository, multiple expert modules and a knowledge manager. The application contains one knowledge miner specific to the domain of FP7 European project proposals. The knowledge miner is implemented by a combination of applying suitable data mining algorithms from PolyAnalyst and manual generation of rules from a generic rule template.

The repository is composed of an EXCEL spreadsheet containing the downloaded project summaries plus internal PolyAnalyst spreadsheets which are created and updated as a result of data mining activities. CMS contains expert modules but these are not currently updated automatically as PolyAnalyst is a standalone application.

The knowledge manager role is performed manually. Further development will enable part of the knowledge manager to be automated but some activities, such as selecting appropriate data to mine and evaluating whether the knowledge discovered is appropriate to store, require human intervention.

5.2. Application Knowledge Discovery Activities

The phases of the knowledge discovery process model are shown in figure 3. The activities that are utilized within the phases are described in the following sub-sections.

5.2.1. Domain understanding

This was achieved was by perusing the ICT work programme document which defines the priorities for calls for funding proposals and the criteria that is used for evaluating the proposals responding to these calls (European Commission, 2007). Existing experience of the authors in preparing FP7 project proposals was also used to provide domain understanding.

5.2.2. Data understanding

The data was located and its existing structure considered. The project summary data was stored as a web page, headings and tables providing a semi-structure.

Each project summary was formatted into the following sections:

- Project title.
- Research area - FP7 research objectives are refined into sub-objectives, e.g. ICT-2007.1.1 Future Internet Architectures and Network Technologies, ICT-2007.1.2 Service and Software Architectures, Infrastructures and Engineering, etc.
- Lead organization details – name, address and other details of the organization coordinating the project.
- Project description – a brief summary of the project's main tasks and goals.
- Project details – including details of duration, cost and type. FP7 projects are classified into types according to funding scheme. The following schemes exist: Collaborative projects; Networks of Excellence; Coordination and support actions.
- Participants – names of the organizations participating in the project.
- Participants' countries.

To maintain the tabulation the data was downloaded as ".htm" files.

5.2.3. Data Preparation

In the data preparation phase the data is structured and processed to extract attributes and derive new attributes to which algorithms may be applied. To structure the data text from the downloaded ".htm" files was copied and formatted into an EXCEL spreadsheet. One project summary was placed on each of the spreadsheet rows and the columns were configured based on the structure described above. Structuring the data enabled non informative words to be removed. For example, the text "Seventh Framework Programme" occurs in every project summary but does not convey any information about why the project proposal was successful. The removal of non informative words reduces document size and avoids information overload. Structuring also gives a context to data. For example, when the word "France" occurs in a column named "Participants' countries" it is obvious that the context relates to a geographical area and not to a man named "Mr. France".

To process the data mining algorithms were applied using PolyAnalyst. In this phase the mining algorithms are utilized specifically to obtain attributes from the data. In the next phase, “application of algorithms”, patterns are located within the attributes and rules are derived from these patterns. To extract attributes simple string matching, keyword analysis and entity extraction were performed. String matching was used to extract project type from the “Project details” column. Keyword analysis applies text analysis algorithms, such as detecting word boundaries, to identify and count frequently occurring keywords and phrases in natural language. Phrases are two or more consecutive words which occur within the same sentence, e.g. “United Kingdom”. Keyword analysis was used to identify individual countries within the “Participants’ countries” column and to extract attributes from the “Lead organization details” column.

Entity extraction identifies a word or a phrase or a pattern of characters occurring within natural language that matches a given structural definition. PolyAnalyst 6.0 provides a set of entities that can be discovered by the inbuilt entity extraction algorithm, e.g. dates and currency amounts. It is also possible to customize entity extraction within PolyAnalyst by creating regular expressions to match selected characters. Entity extraction was employed to extract project cost and duration from the “Project Details” column. Project cost was extracted using the PolyAnalyst currency amount entity. To extract project duration the regular expression "Duration:"(\S)*\#\#” (i.e. match the string “Duration:” followed by optional whitespace and 2 integers) was employed.

New attributes are derived from those extracted. Binning, grouping and context rules were applied to derive new attributes. Binning abstracts continuous data into discrete data by equally distributing numerical values into a defined number of ranges (or bins). Binning involves loss of data but can enhance the meaning. Project cost was binned into “Small Budget”, “Medium Budget” and “Large Budget”. It can be seen that “Small Budget” is more meaningful than, for example, 579,999.00 euro.

Grouping is similar to binning but arranges data into user defined groups according to numerical value. Grouping was used to derive “Number of Rich

Countries", "Number of Poor Countries", "Number of Rich Country Participants" and "Number of Poor Country Participants" within a research project. To find all the countries taking part in a project country attributes extracted from the "Participants' countries" column by keyword analysis were combined with attributes identified as countries in the "Lead organization details" column. Keyword analysis provided a count of the number of countries. The countries were grouped by GDP per capita 2009 (gross domestic product divided by midyear population) (The World Bank) which is a widely used economic benchmark. Locating this additional information required a return to the data understanding phase.

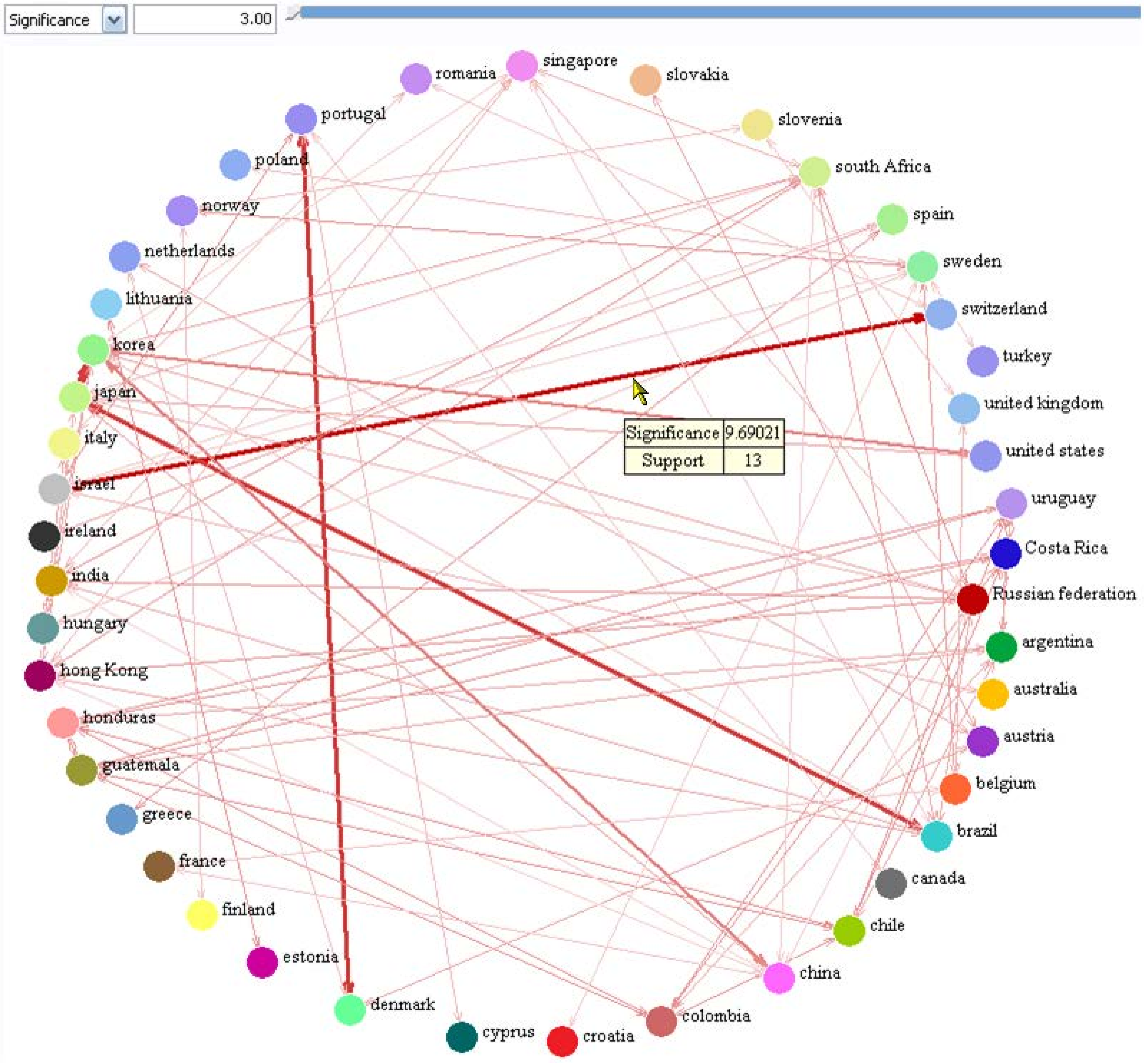
PolyAnalyst 6.0 contains a rule language, Symbolic Rule Language (SRL), for manipulating data and calculating results which can be utilized to create context rules. Context rules assign new attributes that describe the existing information and attributes in terms of general categories, resulting in the generation of new information. Context rules were created to classify the Lead organization into the following Leader Types: Industry, Research and University. A simplified version of one of the context rules is shown below.

```
IF '([GMBH] OR [SAS] OR [plc]) @[Lead Organisation]' OR '"LTD"
@[Lead Organisation]' OR '"Limited" @[Lead Organisation]'
THEN [Industry]
```

"[]" indicates a derived keyword or a spreadsheet column. GmbH is a German abbreviation for Gesellschaft mit beschränkter Haftung and translates to "Company with limited liability." SAS is an Italian abbreviation for Societá in Accomandita Semplice meaning " Limited Partnership". It can be seen that the rule makes use of both string matching and derived keywords. It was not found possible to base the context rule solely on derived keywords as keyword analysis does not pick up infrequent values and to achieve a complete dataset a context was required for every Lead organization.

5.2.4. Application of algorithms

A number of data mining algorithms exist. This research applies Link Analysis, Link Chart and rule generation from a template to the attributes derived during the data preparation phase. Link Analysis highlights patterns of co-occurrence of keywords extracted from data by representing the relations in a graph. Link Analysis was used to consider relationships between all the countries taking part in a project (see “Data preparation” section for a description of how these keywords were derived).



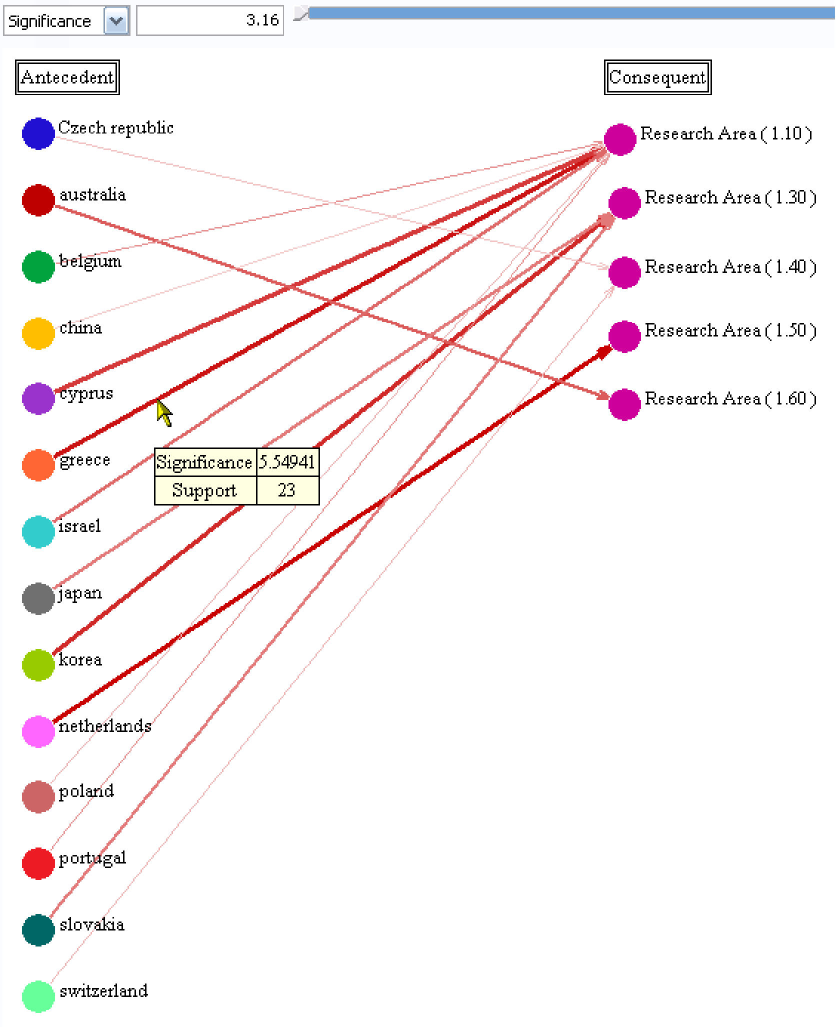
Significance 3.00

Significance	9.69021
Support	13

Figure 4

PolyAnalyst indicates the strength of a relationship by the thickness of the line linking the keywords. It can be seen from figure 4 that there is a strong relationship between Israel and Switzerland, i.e. in projects in which Israel part takes there is often a Swiss participant. Co-occurrence is measured by support and significance. Support is the number of records (spreadsheet rows) in which the two related attributes occur. Significance is the absolute value of the logarithm of probability that the relationship occurred by chance and is determined by the binomial distribution. PolyAnalyst stores support, significance and the record sets supporting the relationships in underlying tables which can be accessed by clicking on the Link Analysis graph.

A Link Chart presents a more structured version of Link Analysis and considers relationships between two specified dataset columns configured as antecedent and consequent columns. An example of a Link Chart is shown below. To perform Link Chart analysis the keywords need to be placed into sets or structured as attributes with sets of predefined values. Grouping or binning techniques can be employed to define sets of attribute values (see sub-section 5.2.3 "Data preparation"). For example, to create the Link Chart shown below , "Countries Participating in a Project vs. Research Area", all the keywords describing countries were placed in a set. The attribute "Research Area" can be assigned any value from the pre-existing set [1.1 – 1.6]. To detect "hidden" relationships in the data, sets of extracted attributes were systematically paired and their values subjected to Link Chart analysis.



Significance	5.54941
Support	23

Figure 5

To perform rule generation a template has been derived which enables rules to be semi-automatically discovered. A rule is generated by inserting the attributes of a statistically significant relationship detected by Link Chart into the template. Statistical significance is determined in the evaluation of results phase and is explained in the next section. To discover rules the following generic rule template is used:

IF Antecedent is X and Consequent is != Y
THEN send message "Consider if Consequent is suitable"

Antecedent is the name of the Link Chart antecedent column and X is a value from this column. Consequent the name of the Link Chart consequent column and Y is the consequent column value which correlates with X. For example, selecting the antecedent value "Greece" from the Link Chart shown in figure 5 and substituting it along with its related value into the rule template will result in the following rule.

IF Country is Greece and Research Area != 1.1
THEN send message “Consider if Research Area is suitable”

It can be seen that several other rules can be generated from this Link Chart. Rules with the same consequent value are combined to remove superfluous activations. The rule shown below could also be generated from the Link Chart shown in figure 6.

IF Country is Cyprus and Research Area != 1.1
THEN send message “Consider if Research Area is suitable”

This is combined with the previous rule to give:

IF Country is Greece OR Country is Cyprus and Research Area != 1.1
THEN send message “Consider if Research Area is suitable”

Currently rule generation is performed manually but this part of the rule discovery process could be readily automated. Additionally, for Link Chart or Link Analysis relationships for which no meaningful rules can be generated from the template, the expert may choose to create rules. Examples of expert defined rules are shown in the Results section.

5.2.5. Evaluation of Results

This phase assesses if the relationships discovered through Link Analysis and Link Chart are statistically significant and whether rules generated from the generic template contain new information and are applicable within the domain. This research applies a significance level of 5%, which is commonly used in non–critical applications. PolyAnalyst 6.0 records significance as the negative antilogarithm of the value, hence all values above 3 are regarded as statistically significant (antilog (-3) = 0.05). At the moment, checking the discovered relationships for statistical significance is a manual process but this would be simple to automate.

The rules generated from the generic template together with their supporting data are displayed to an expert for verification. The expert checks that the information contained by the rules is novel and does not duplicate existing information within CMS KB by using a different format or terminology. The expert might decide to change the text of a rule's message to clarify meaning or equate with natural speech. For example, to aid user understanding, the expert might chose to alter the rule shown in the "Application of algorithms" section as follows:

IF Country is Greece and Research Area != 1.1
THEN send message "Consider if Participant has sufficient experience in this Research Area"

The expert may also delete or modify a rule deemed to be incorrect. Although the rules are derived from statistically significant relationships, statistical correlation does not infer causality which is implied by the rules. For example, the following rule could be generated from the values of the attribute relationships shown in table 1:

IF Country is Austria and Project Cost != Large Budget
THEN send message "Consider if Project Cost is suitable"

This rule implies that members located in Austria prefer a project with a large budget and that this will form one of the contributory causes in the project proposal being successfully funded. This does not appear to be a good rule and the underlying reasons for the relationship upon which this rule is based are unclear, therefore this rule is deleted by the expert. Hence, it can be seen that there is a need for the expert to verify that the generated rules are valid.

5.2.6. Deployment and Conflict Resolution

As PolyAnalyst is a standalone application these stages have not been implemented. During the deployment stage, when future development has integrated data mining algorithms into CMS, relevant expert modules will be updated through the Enterprise Service Bus. The Enterprise Service Bus connects the CMS Knowledge Acquisition Module to CMS KB containing the expert modules (see section 3 "Collaboration Moderator Services Structure").

Rules deployed to the expert modules might conflict with existing rules contained by the expert modules which would be based on data mining of previous data sources. If potential project consortia choose not to (or are unable to) complete the advice offered by CMS but are still successful, new rules will be generated which could conflict with existing rules. The new rules may consider a broader, narrower or different set of relationships to the existing rules. For example, the rule given in the "Application of algorithms" section might be broadened to

> IF Country is Greece OR Country is Cyprus OR Country is UK and
> Research Area != 1.1
> THEN send message "Consider if Research Area is suitable"

if more projects with UK members became successful in research area 1.1. If Cyprus was found to be no longer successful in this research area the rule would consider a narrower attribute set. A domain expert is needed to decide whether existing rules in conflict with new rules should be updated, deleted or amalgamated with the new rules. Conflict resolution enables CMS KB to dynamically evolve.

The deployed rules could be utilized by CMS to support future project teams (VOs) in preparing a European research project proposal. CMS could use the rules to alert team members to factors that had been important in other successful proposals in the past.

6. Results

Table 1 shows the results achieved by applying the Knowledge Discovery Framework described in section 4 to project summaries of successful past European FP7 research projects. The table summarizes the relationships derived when Link Analysis and Link Chart algorithms were applied to attributes extracted by the data mining techniques described in the "Data Preparation" section. The table also shows the rules semi-automatically generated from these relationships. For brevity, not all the relationships and rules found are shown. Although Link Chart algorithms were used to explore relationships between all of the extracted attributes only statistically significant relationships are shown.

Table 1 presents the names and subsets of the values of Link Chart antecedent and consequent attributes and the statistical significance level of the relationships. The "Lead Country" attribute is the country where the lead organization of the research project is located. The “Country” attribute defines all the countries taking part in a project and is derived from merging keyword analysis of the “Participants’ countries” column with country attributes identified within the “Lead organization details” column. For explanations of the other attributes see sub-sections “Data understanding” and “Data preparation”.

For reasons of space only one rule generated from a generic template is provided for each antecedent/ consequent pair, based on the subset of values given in the table, and the action “send message” is left out from the template as this can be assumed. Only rules which have been verified by an expert are given. Where appropriate, a version of the rule which has been modified by the expert is provided. Rules which have been derived manually by the expert based on the discovered relationships are also given.

<<Insert Table 1 about here>>

7. Findings from the Application

The results demonstrate that it is possible to utilize the knowledge acquired from data mining to semi-automatically generate rules from a generic template. This section considers what has been discovered about this approach.

It was found that, as many researchers note (Zhang et al, 2003; Mlynarski et al, 2006; Alcala-Fdez et al, 2009), the most time consuming activity implemented by the Knowledge Discovery Process model was the Data Preparation phase (see section “Knowledge Discovery Framework”). Both structuring the data into a spreadsheet and processing the data by applying mining algorithms to extract attributes were found to be lengthy activities. To maintain the semi-structure provided by the original web page format the data needed to be manually copied into the spreadsheet. Whilst PolyAnalyst provides some algorithms enabling the automatic derivation of attributes, such as entity extraction, other algorithms need user input in the form of regular expressions or rules which can be complex to create. To apply data mining algorithms in the next phase of the knowledge discovery process, the user first needs to group and filter the derived attributes. In order to generate meaningful rules it was discovered that the consequent attribute (see section “Application of Algorithms”) must have a small number of possible values. For example, considering the following rule taken from the results in Table 1:

IF Lead Organization is University L and Research Area!= 1.4
THEN send message “Consider if Research Area is suitable”

For this rule the antecedent attribute “Lead Organization” has a wide range of possible values (140 organizations were present in the data source). The consequent attribute “Research Area” can take one of six possible values from the set [1.1- 1.6]. If the antecedent and consequent attributes were to be reversed the new consequent attribute “Lead Organization” not have a small number of possible values and the rule would become:

IF Research Area is 1.4 and Lead Organization != University L
THEN send message “Consider if Lead Organization is suitable”

Obviously the rearranged rule is not sensible. Looking at Table 1, it can be seen that no rule could be derived for the attribute pairs Country/ Country and Lead Country/ Country as for these attribute combinations the consequent attribute cannot have a small number of possible values. (There were 19 Lead Countries and 51 Countries present in the data source). From the results achieved by this research, exactly how small the number of possible consequent attribute values is required to be to generate meaningful rules cannot be determined but it appears to be less than ten.

Where the number of possible values of both the antecedent and consequent attributes is small, the expert must decide which combination produces the most sensible rules. For example, considering the rule below taken from Table 1.

IF Research Area is 1.3 and Project Cost != Small Budget
THEN send message “Consider if Project Cost is suitable”

Reversing the antecedent and consequent attributes would result in:

IF Project Cost is Small Budget and Research Area is 1.3
THEN send message “Consider if Research Area is suitable”

It can be seen that the original rule is more appropriate. The example above suggests that where the number of possible values for both the antecedent and consequent attribute is small the consequent attribute should have a smaller number of possible values than the antecedent. ("Research Area" has six possible values, "Project Cost" can take one of three possible values from the set [Small Budget, Medium Budget, Large Budget]). However, as the example shown was the only rule of this type generated no conclusions can be drawn.

It was also found that it was not possible to generate meaningful rules from the generic template for attributes with values from the set of continuous numbers, although the expert was able to create rules for relationships containing this type of attribute. For example, no rules could be generated by substituting values from the attribute relationship Number of Rich Countries/ Number of Poor Countries into the generic template although the expert was able to create a rule for this relationship, as shown in Table 1.

To generate rules from the generic template the Link Charts which detect the relationships used to generate the rules must contain datasets of suitable attributes which are appropriately ordered. If a Link Chart algorithm was to be incorporated into CMS, the selection and ordering of appropriate attribute combinations for Link Charts could be automated.

It was possible to generate rules from a generic template for ten of the 15 suitable attribute relationships. In addition, the expert was able to create rules for three more attribute relationships, resulting in rules for 13 of the 20 statistically significant attribute relationships discovered by data mining the European research project summaries For most of the rules generated the expert changed the rule's message text to clarify meaning. For some of the attribute relationships no sensible rules could be generated (see Table 1). The attribute relationships were found by systematically pairing all the attributes extracted during the data preparation phase (see sub-section 5.2.3) and subjecting the attribute values to Link Chart analysis. To reduce the number of spurious rules generated and the need for the expert to check and delete them, more guidance is required as to which attribute pair combinations are appropriate. An overall evaluation of the

SYNERGY system, of which this research forms a part, is presented in Lorré, et al (2011).

8. Conclusions

This paper has demonstrated the implementation of a knowledge discovery framework which utilizes information acquired from data mining techniques to semi-automatically generate rules for a Moderator. The results achieved demonstrate the feasibility of this approach. The rules are generated by substituting the relationships derived from data mining into a generic template. The resulting rules need to be verified by an expert to ensure that the information they contain is novel, understandable and applicable within the domain of interest. The expert may also chose to create rules based on the patterns and relations discovered as a result of data mining. The combination of automated rule generation from a generic template with expert verification and expert derived rules enables knowledge to be mined semi-automatically reducing the cost of development by decreasing the amount of time needed from experts and knowledge engineers. The authors believe that the approach presented in this paper, which provides a semi-automatic methodology to generate rules by inserting relationships discovered as a result of data mining into a generic template, is novel.

The approach is demonstrated by an application case which focuses on the VO Pre-creation stage. The rules generated by implementing the knowledge discovery process model can be utilized by CMS to support future project teams (VOs) in preparing a European project proposal. CMS could use the rules to make team members aware of factors that had been important in other successful proposals in the past. To achieve results within a limited time frame, the application case considers FP7 Challenge 1 European project summaries but it could be applied to the complete FP7 project database. An application case could also be developed to extend the work of Grobelnik and Mladenic (2003) by utilizing the relationships they discovered to generate rules.

To assess the feasibility of the approach the application utilized the standalone application PolyAnalyst 6.0. Future development is required to assess suitable open source data mining tools to incorporate within CMS to achieve the functionality of the procedures discovered using PolyAnalyst. The current rule generation method and the method evaluating rules for statistical significance require automating. The selection and ordering of appropriate attribute combinations for Link Charts needs to be automated. When these developments have been achieved an evaluation of the reduction of expert time required for knowledge acquisition for a Moderator will be possible.

Future research needs to consider two aspects: using an ontology and the development of more generic rule templates. The adoption of an ontology, such as ECOS (see section 3), by the knowledge discovery framework would provide a shared understanding of terms between the CMS Knowledge Miner for Learning, the knowledge manager and CMS KB. Knowledge discovery would be based upon the common ontology, so that the knowledge generated and used to update the expert modules contained in the CMS KB would be consistent and structured in an appropriate format. The use of an ontology would reduce the amount of work required by the expert in evaluating the information contained by the rules generated. The expert would not have to check whether the information contained by the rules duplicates existing information within CMS KB by using a different format or terminology. The ontology could be used to provide guidance on suitable attributes to derive during the data preparation phase and to suggest appropriate attribute pair combinations for Link Chart analysis. This would decrease the number of spurious rules generated and hence the number of rules that the expert is required to check and delete. For more details on the potential use of an ontology see Palmer, C. (2011), Applying an Enterprise Ontology to Guide Data Mining, Natural Computing Application Forum meeting, July 2011, Swansea University, U.K.

This research has derived one generic template which enables rules to be semi-automatically discovered for a Moderator. Further work is needed to determine if more templates can be developed (and are needed). The template is limited to considering one conjunctive (“And”) relationship between attributes (see the

“Results” section for examples). This is because the template is intended to generate rules from the results of Link Chart analysis which considers relationships between two columns of attributes. More research is required to consider whether it would be possible to develop generic templates containing no conjunctions or more than one conjunctive relationship between attributes from suitable data mining algorithms. To derive more templates would require the application of more data mining algorithms and an assessment of the results.

To summarize the general findings of this research:

- The most time consuming activity in the process of discovering knowledge from data was the Data Preparation phase.
- Rules for a knowledge-based system may be generated by inserting relationships discovered as a result of data mining into a generic template.
- To generate meaningful rules the consequent attribute of a discovered relationship must have a small number of possible values (less than ten).
- More guidance is required as to which attribute pair combinations are appropriate to discover relationships from which rules can be generated. This guidance could be provided by an ontology.
- The rules generated need to be verified by an expert, resulting in a semi-automatic methodology for knowledge discovery.

ACKNOWLEDGEMENTS

This research is part of the SYNERGY project, under European Commission, 7th Framework Programme Grant ICT-2007.1.3-216089.

REFERENCES

Alcalá-Fdez, J., Sánchez, L., García, S., del Jesus, M., Ventura, S., Garrell, J., Otero, J., Romero, C., Bacardit, J., Rivas, V., Fernández, J., Herrera, F. (2009). KEEL: a software tool to assess evolutionary algorithms for data mining problems, Soft Computing - A Fusion of Foundations, Methodologies and Applications, 13(3), 307-318.

Choudhary, A.K., Harding, J.A., Lin, H. K., Tiwari, M. K. and Swarnkar, R . (In press). Knowledge discOvery And daTa minINg inteGrated (KOATING) Moderators For Collaborative Projects. Manuscript submitted to International Journal of Production Research.

Choudhary, A.K., Harding, J.A. and Tiwari M.K. (2009). Data mining in Manufacturing: a review based on the kind of knowledge, J Intell Manuf, 20, 501- 521.

Choudhary, A.K. , Oluikpe, P.I., Harding J.A. and Carillo P.M. (2009a). The needs and benefits of Text Mining applications on Post-Project Reviews, Computers in Industry, 60, 728-740.

European Commission (2007) ICT – Information and Communication Technologies - A Theme for research and development under the specific programme "Cooperation" implementing the Seventh Framework Programme (2007-2013) of the European Community for research, technological development and demonstration activities -Work Programme 2007-08. ftp://ftp.cordis.europa.eu/pub/fp7/ict/docs/ict-wp-2007-08_en.pdf. Accessed 25 February 2011.

Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases, AI Magazine, 17 (3), 38 – 54.

Gasar, S., Bohanec, M. and Rajkovic, V. (2003). A Combined Data Mining and Decision Support Approach to Educational Planning. In D. Mladenic, N. Lavrac, M. Bohanec and S. Moyle (eds), Data Mining and Decision Support Integration and Collaboration, (pp. 203-212). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Grobelnik, M., and Mladenic, D. (2003). Analysis of a Database of Research Projects using Text Mining and Link Analysis. In D. Mladenic, N. Lavrac, M. Bohanec and S. Moyle (eds), Data Mining and Decision Support Integration and Collaboration (pp. 157-166). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Han, J. and Kamber, M. (2006). Data Mining Concepts and Techniques (2nd ed.), Morgan Kaufmann, San Francisco, CA, USA.

Harding, J.A., Shahbaz, M., Srinivas, S, and Kusiak, A. (2006). Data Mining in Manufacturing, Transactions of the American Society of Mechanical Engineers (ASME): Journal of Manufacturing Science and Engineering, 128, 969 - 976.

Harding, J.A., Popplewell, K. and Lin, H.K. (2007). A Generation of Moderators from Single Product to Global e-Supply Chain, In Knowledge and Technology Management in Virtual Organizations: Issues, Trends, Opportunities and Solutions (Chapter V, pp,110-135), G.D. Putnik and M.M. Cunha, eds, Idea-Group Publishing, USA, ISBN 159904166-9.

Harding, J.A., Popplewell, K. and Cook, D. (2003). Manufacturing system engineering moderator: an aid for multidiscipline project teams, International Journal of Production Research, 41(9), 1973.

Heath, J. (2006). A framework for an Intelligent Decision Support System (IDSS), including a data mining methodology, for fetal-maternal clinical practice and research. Thesis (MSc). University of Western Sydney, Australia. Available from: http://handle.uws.edu.au:8081/1959.7/14810

Lorré, J-P., and Rajsiri, N. (2011). D7.3: Evaluated SYNERGY System, EBM Websourcing, France, http://www.synergy-ist.eu/.

Khilwani, N. (2011). Role of Semantic Web in Changing Context of Enterprise Collaboration, Thesis (PhD). Loughborough University, UK.

Khilwani, N., Harding, J A, Tiwari, M K, (2011), "ECOS: Publishing the Published Competences" , *Proc. IMechE, Part B: J. Engineering Manufacture*, volume 225 issue 6, pages 921-942, doi:10.1177/09544054JEM2097, ISSN 0954-4054,

Lima, C., Costa, R , Malo, P. and Antunes, J. (2010). A Knowledge-Based Approach to Support Decision Making Process in Project-Oriented Collaboration, Proceedings of the 11th European Conference of Knowledge Management, 1-2, 614 -622.

Lin, H. K., Harding, J. A. and Teoh, P. C. (2005). An inter-enterprise semantic web system to support information autonomy and conflict moderation, Proc. Inst. Mech. Eng. Pt. B: J. Eng. Manuf., 219, 903-911.

Lavrac, N. and Bohanec, M. (2003). Integration of data mining and decision support. In D. Mladenic, N. Lavrac, M. Bohanec and S. Moyle (eds), Data Mining and Decision Support Integration and Collaboration(pp. 36-480, Dordrecht, The Netherlands: Kluwer Academic Publishers.

Luczak, H and Hauser, A. (2005). Knowledge management in virtual organizations, Proceedings of ICSSSM'05, International Conference on Services Systems and Services Management, 898.

Mezga´r, I. and Kova´cs, G. L. (1998). Co-ordination of SME production through a co-operative network. J Intell Manuf, 9 (2), 167- 172.

Mladenic, D., Lavrac, N., Bohanec, M. and Moyle, S.,eds, (2003). Data Mining and Decision Support Integration and Collaboration, Dordrecht, The Netherlands: Kluwer Academic Publishers.

Mlynarski, R., Ilczuk, G., Wakulicz-Deja, A. and Kargul, W. (2006). A new method of data preparation for cardiological decision support. Computers in Cardiology, Sept, 273-276.

Nof, S.Y. (2006). Collaborative e-work and e-manufacturing: challenges for production and logistics managers, J Intell Manuf , 17 (6), 689–701.

Petalslink.com. Open Source enterprise services bus (ESB). http://www.petalslink.com/en/products/petals-esb. Accessed 25 February 2011.

Polyanalyst 6.0. www.megaputer.com. Accessed 25 February 2011.

Shearer, C. (2000) The CRISP-DM model: the new blueprint for data mining, Journal of Data Warehousing, 5, 13-22.

Swarnkar, R., Choudhary, A., Harding, J., Das, B. and Young, R. (2011). A framework for collaboration moderator services to support knowledge based collaboration, J Intell Manuf, DOI 10.1007/s10845-011-0528-2m published online.

The World Bank, GDP per capita (current US$). http://data.worldbank.org/indicator/NY.GDP.PCAP.CD .Accessed 25 February 2011.

Wang, K. (2007). Applying data mining to manufacturing: the nature and implications. J Intell Manuf , 18 (4), 487–495.

Zhang, S., Zhang, C. and Yang, Q. (2003) Data Preparation for Data Mining, Applied Artificial Intelligence, 17, 375–381.

Tables

Table 1 Results Derived from Link Analysis and Link Charts

List of Figures

Fig. 1 CMS modules and repositories

Fig. 2 Knowledge Discovery Framework For CMS

Fig. 3 Knowledge Discovery Process Model

Fig. 4 Link Analysis of Countries Participating in a Project

Fig. 5 Link Analysis of Countries Participating in a Project vs. Research Area