Knowledge Discovery from Post-Project Reviews

Abstract

Many construction companies conduct reviews on project completion to enhance learning and to fulfil quality management procedures. Often these reports are filed away never to be seen again. This means that potentially important knowledge that may assist other project teams is not exploited. This paper investigates whether Knowledge Discovery from Text (KDT) and text mining (TM) could be used to "discover" useful knowledge from such reports. Text mining avoids the need to manually search a vast number of reports, potentially of different formats and foci, to seek trends that may be useful for current and future projects. Pilot tests were used to analyse 48 post-project review reports. The reports were first reviewed manually to identify key themes. They were then analysed using text mining software to investigate whether text mining could identify trends and uncover useful knowledge from the reports. Pilot tests succeeded in finding common occurrences across different projects that were previously unknown. Text mining could provide a potential solution and would aid project teams to learn from previous projects. However, a lot of work is currently required before the text mining tests are conducted and the results need to be examined carefully by those with domain knowledge to validate the results obtained.

Keywords: knowledge discovery, text mining, project reviews, knowledge.

Introduction

Organisations of all descriptions are trying to become more competitive. This means examining their processes and trying to find those with scope for improvement. In the construction sector, Fairclough (2002) in his report on improving the competiveness of the UK construction industry's innovation and research, stated "there is a problem with institutional learning to capture this innovation for future projects". He also lamented that "with a few notable exceptions, knowledge transfer does not tend to 'ripple' out from members of project teams to their companies or other organisations". One way of overcoming this problem is to learn from previous projects. Whilst there are many mechanisms to do so, Post-Project Reviews offer a possible solution since they attempt, to differing degrees of detail, to document what went well and where there is scope for improvement. However, although many organisations acknowledge that such reviews can provide useful project knowledge, there are a host of issues that serve to undermine the importance of such reviews. Text mining is one possible solution to coping with the vast amount of untapped knowledge stored within these project reviews.

This paper commences by investigating the importance of organisational learning and the use of Post-Project Reviews as one mechanism to promote learning from construction projects. It then focuses on one tool - text mining- as a possible mechanism to learn from Post-Project Reviews. A number of Post-Project Reviews from two companies are analysed in an attempt to extract the largely tacit knowledge that can be used to improve future projects. The challenges of using text mining tools on Post-Project Reviews are identified and the quality of the results obtained is critiqued. The paper concludes with a discussion on the results of the experiments conducted, the limitations of the method adopted and the implications for future research.

Learning and Post-Project Reviews

Organisations are keen to improve on processes with the hope that they will generate better performance. For such organisations, there is plenty of advice on how they may learn in order to improve their performance. The field of organisational learning has been investigated for decades. Arygris (1977), based on his work to understand why organisations do not learn from previous mistakes, introduced the concept of double loop learning. He provided a definition of organisational learning as "a process of detecting and correcting error". Since then many authors have written on the subject. More recently Easterby-Smith et al. (2000) provided a useful time-line of the key debates on organisational learning. They identified that the key debate between individual and organisational levels had subsided and the role of group level learning had taken a more prominent role. Current debate revolves around three main issues: (1) the nature and location of organizational learning; (2) how to investigate organisational learning; and (3) territorial debates. Within the context of this paper, the latter is important. Easterby-Smith et al. identified the tensions between organisational learning and knowledge management. However, they noted that there was growing convergence between both communities as they had looking at similar concepts and problems albeit using different terminologies.

Construction organisations have embraced the concept of continuous improvement and leaning from experience, most notably through their use of "best practice". One mechanism for promoting the concept of organisational learning is to investigate what happens on projects. There are different types of project reviews and these may take place at different stages in the project lifecycle. However, Post-Project Previews are particularly appealing since they document the full history of the project, and not just parts of the project. They play a fundamental role in promoting organisational learning. von Zedtwitz (2002) conducted research over a five-year period with Research and Development managers of leading companies. He concluded that PPRs should not be conducted for their own sake and any

outcome must be an input into future projects. This resonates with the aim of finding useful knowledge from Post-Project Reviews that can contribute to learning for future projects.

Post-Project Reviews (PPRs) are potentially a rich source of knowledge for organisations. von Zedtwitz (2002) defined PPRs as “a formal review of the project examining the lessons which may be learned and used to the benefit of future projects”. It is a process through which a project team looks at the project outcome retrospectively with a view to learning from activities carried out, to avoid mistakes and also to learn from successes and failures. PPRs have been around for over 40 years when they were initially used by the US. Since then the concept has been adopted in many organisations under a number of different names. Disterer (2002) investigated the transfer of project knowledge and experiences. He summarised the various terms used by authors to represent an opportunity to identify and secure knowledge and experiences of project team members. These include Post-Project Reviews, post-project appraisals, project post-mortem, debriefing, reuse planning, reflection, corporate feedback cycle, experience factory, etc. Roth and Kleiner (1998), in a n effort to improve organisationlearning, also introduced the term “learning histories”. These allow organisations to reflect on past experience leading to effective future actions. In the UK, case studies show that major companies such as BP Amoco, BAA plc, National Grid, Transco and construction companies such as Bovis Lend Lease, and Buro Happold have adopted this methodology in an effort to learn from experience (David Bartholomew Associates, 2003).

There is a lot of guidance on how to conduct PPRs. Some of these include Baird et al. (1999) which advocates a phased approach to PPR although it is unclear on details. Sowards (2005) recommended a five stage process which focused on establishing criteria, involving key people, discussing an agenda, documenting key learning points and disseminating these to people who should see them. Roth and Kleiner (1998) suggested a six stage process which begins with planning, reflective interviews, distillation, writing,

validation and dissemination. Busby (1999) advocates a process that integrates feedback loops for effective learning but came short of elaborating on how this could be done. Schindler and Eppler (2003) suggest that a "project knowledge broker" should be responsible for the processes of reviewing a project and transferring lessons learnt within and between project teams. This recommendation buys more into the idea of outsourcing the review process and this would have obvious implications for the organisation. With regard to PPRs specific to the construction sector, the US's Construction Industry Institute (2007) provides practical guidance and Gibson et al. (2007) provide a useful critique of existing post-project view practices.

Benefits of conducting PPR

The research done in this area supports the numerous benefits of conducting Post-Project Reviews. These may be summarised as follows:

- Facilitating collective learning (Kerth, 2000, Tan et al., 2006). PPRs allow project teams to come together to discuss and debate project issues and thereby obtain another perspective;
- Prevents knowledge loss (Carrillo, 2005). When project teams disband to work on other projects, individual knowledge can be loss. PPR reports provide an opportunity to record that knowledge for future use;
- Promotes continuous improvement (Prahalad and Hamel, 1990). This can benefit client organisation if evidence of PPRs provide confidence that construction companies are actually reviewing the performance of projects with a view to enhancing future projects.

Thus several benefits can accrue from conducting PPRs. The real challenge is to use PPRs to their full potential given the difficulties outlined in the next section.

Difficulties with PPR

Whilst many authors have praised the benefits accrued from PPRs. A number of practical issues have also been identified. With respect to the construction industry, these include the following:

- Ad hoc processes - although organisations may have procedures for conducting project reviews, these may not be systematic and followed throughout the organisation;
- Availability of key staff - the nature of construction projects mean that key staff may be transferred to other projects and thus may not be available for conducting such reviews;
- Timing - this is a key issue with regard to when these are held. Many companies prefer to wait until all project expenses have been accounted for by which time key staff have other responsibilities (Carrillo, 2004); and
- Content - Fairclough (2002) pointed out the mistrust of such lessons as a key inhibitor. The PPR must have a useable format, length and content.
- Dissemination - even though organisations may have review reports, the dissemination of key knowledge and lessons are not as expected

The above challenges are not insurmountable but require resources that are not always readily available. Rather than leave these potentially useful reports untapped, text mining will be explored to gauge whether it can help to discover knowledge from PPRs.

Knowledge Discovery and Text Mining

Hotho et al. (2005) described how text mining or Knowledge Discovery from Text (KDT) was first mentioned by Feldman and Dagan (1995). Later, Feldman and Sangar (2007) defined

text mining as “the process of discovering information in large text collections, and automatically identifying interesting patterns and relationships in textual data”. Text mining is a natural solution for transforming unstructured information stored in numerous text documents into useful knowledge. The extracted knowledge can be used to model, classify and make predictions for numerous applications. Text mining is sometimes considered to be a subset of data mining, focused on the analysis of textual rather than numerical data. Data Mining has been adopted in many business sectors, including manufacturing, during the 1990s and has been gaining importance since 2000 (Harding et al. 2006). Data mining relies on the analysis of numerical data that is stored in a structured manner, typically in large databases. Text mining uses unstructured textual data from a number of different sources. Hearst (2003) points out text mining is different from normal searches where the user is typically looking for something that is already known and has been written by someone else. In text mining, the user attempts to discover knowledge that no one yet knows and so could not have yet written down. The use of text mining is relatively new (Feldman and Sanger, 2007), although it is arguable that it originates from a paper published by Luhn (1958) in which he presented the idea of automatic abstracting text based on the words which are significant in the document.

Text mining evolved from Knowledge Discovery in Databases (KDD) and uses the same techniques for information retrieval, information extraction as well as natural language processing (NLP) and connects them to the algorithms and methods of KDD. Hotho et al (2005) point out that instead of numerical data, text documents are the focus of the analysis. They also point out that these steps have to be applied to a data set in order to extract useful patterns. These steps have to be performed iteratively and several steps usually require interactive feedback from a user. The Cross Industry Standard Process for Data Mining (CRISP) (2010) defines the main steps as (1) business understanding, (2) data understanding, (3) data preparation, (4) modelling, (5) evaluation, and (6) deployment. KDD is therefore an iterative process and the text cleaning and integration, text selection and

transformation are also very important steps as care must be taken to apply different text mining algorithms in an effective and appropriate manner to produce useful results. Indeed the data cleaning and selection steps generally require more time and effort than the actual text mining. The field of text mining is attracting increasing attention.

Research thus far has focused on improving the technical aspects of text mining with little done on how the results can be exploited. For example, Nahm and Mooney (2003) demonstrated that the knowledge discovered from an automatically extracted database is close in accuracy to the knowledge discovered from a manually constructed database. The few articles documenting the use of text mining for engineering problems include Kasravi (2004) and Huang and Murphey (2006). Kasravi described the extraction of knowledge for downstream engineering applications from vast amounts of information held in various text formats such as books, conference papers, product catalogs and web pages. Huang and Murphey used text mining in the automotive industry to link problem description to diagnosis categories. In the construction field, Caldas et al. (2002) and Caldas and Soibelman (2003) used text mining as part of their methodology to classify construction project information. These articles demonstrated how text mining could automate processes but they were not used to discover "new" knowledge. However, their case study involving 16 project teams and a database of around 4,000 document files of various formats resonate with need to identify useful knowledge from numerous construction documents such as Post-Project Review reports.

Research Method

An experimental methodology was adopted. Fellows and Liu (2003) and Hicks (1982) described an experimental methodology as one where tests are conducted to investigate relationships between activities carried out and the resultant outcomes. This methodology was also successfully used by previous studies of text mining in the construction domain

(Caldas et al, 2002; Caldas and Soibelman, 2003). In this research, experiments were conducted using a software tool to identify whether there were certain trends occurring over a number of PPR reports that could provide useful knowledge for future construction projects. The results obtained from the experiments were then evaluated by the companies providing the reports to investigate their usefulness.

The reports were obtained from two construction companies; these companies were selected because they were industry partners on the research project. Company A is a large contractor organisation providing building, civil engineering and maintenance services across the whole life of projects. Company B is a medium sized architectural and construction company that works with financial, property and retailing companies. The research method was divided into five discrete steps as explained below.

Step 1 involved a review of the PPR documentation provided by the two companies. In all 48 PPR reports were obtained (27 from Company A and 11 from Company B); this allowed the research team to understand the scope and format of the PPR conducted by the two companies. Typically each PPR report consisted of 8 to 20 pages.

Step 2 comprised interviews with key personnel in order to gain an understanding of each company's approach to PPRs. This corresponds with the Cross Industry Standard Process for Data Mining (2010) step 1 on business understanding. This allowed the research team to understand each company's PPR process, specifically the aim of conducting PPRs, when they were held, the key participants in the review process, the agenda and format of the reviews, the supporting documents used, the documentation and dissemination of the PPR meetings. This resulted in a process map of each company's PPR processes. The interviews were also used to extract information from each company on what they considered to be *key knowledge areas* that they thought would be useful to them. This allowed the research team to understand each company's critical success factors in terms of learning from completed projects. These *key knowledge areas* also gave an indication of the topics that should be investigated as part of the text mining process. Tables 1 and 2 show the *key knowledge areas* initially identified by the two companies. Company A provided a

hierarchy based on seven high-level knowledge areas with sub categories. Company B found this difficult to do and provided a broader grouping based on four categories. This then led to the development of hierarchies of keywords / phases for each company which were likely to occur within the PPR reports. These hierarchies were then used in the text mining pilot tests.

<*Insert Table 1 here*>

<*Insert Table 2 here*>

Step 3 explored the data/text mining software tool that should be used. Five data mining tools were selected, investigated and evaluated against various criteria which include flexibility, ease of use, cost, scalability, performance, file size constraints and text mining capacity.

The data mining tools evaluated were:

- BM DB2 Data Warehouse;
- TextAnalyst;
- PolyAnalyst;
- Clementine; and
- SAS Enterprise Miner.

The products were also selected on their perceived ability to meet the needs of the research, popularity, features highlighted in the product description and reviews from other users. Based on these criteria, the PolyAnalyst data mining tool was selected for this project for a number of reasons. The software was easy to use and provided all functionality and performance needed for mining textual data from typical PPR reports, the cost of PolyAnalyst was modest and the after sale service and conditions were favourable when compared to other software tools.

Step 4 consisted of the text mining pilot tests. Two batches of reports were experimented on; one from each company to provide individual company results. The pilot tests consisted of a number of sequential stages including (1) the Preparation Stage, (2) the Experiments Stage, and (3) the Results Analysis Stage.

The *Preparation Stage* consisted of pre-formatting reports which removed colours, lines, boxes, photographs, etc. and converted documents from MS Word to MS Excel formats. The

reports were also manually examined to manually select keywords and phrases that would be used to determine the success of the pilot tests.

The *Experiments Stage* consisted of pre-processing the reports and applying algorithms. The pre-processing stage removed "unwanted" text from the analysis that did not directly impact on keywords or project outcome e.g. pay, wage, bonus, etc. It also identified synonyms to remove duplication in the results. Examples were "KPI" and "Key Performance Indicator" and "sub contractor" and "sub-contractor". Next the algorithms were applied. These consisted of the semi-automatic *Text Analysis* which identified frequently occurring keywords and phrases. This identifies frequency but its limitation is that it may not be a good indicator of relevance or importance. The thematic context of the PolyAnalyst software was then set to cover the key areas of the organisations' business areas such as "business", "science", "technology", etc. Rules Application was used to identify all words belonging to a particular category e.g. if the keyword "finance" is used , all words and phrases that relate to finance such as "additional cost", "budget", "contract sum", "profit" are identified. The rules applied to the dataset allow a statistical analysis to be completed based on the frequency of certain words. Next, a technique called *Link Analysis* was used. Link Analysis is a process through which relationships between words and phrases in the body of the reports were explored. The strength of the relationship is normally indicated by how thick the lines linking the words are. The basic principle underlying Link Analysis is the statistical principle of correlation which explores the strength of relationships between variables and whether these are significant or not. For example, "plan" was found to be linked to "extension of time" in some reports. Lastly, a technique called *Dimensional Matrix Analysis* was used. Dimensional Matrix Analysis compares a number of keywords in the reports and investigates their influence on each other. A Dimensional Matrix was created with columns representing each high-level knowledge area with several key words and rules (e.g. seven for Company A and four for Company B, see Table 1). Each column consisted of different cells where each cell represented the keyword(s) to be searched for within the PPR reports. The user defines

the values of one or more cells and PolyAnalyst browses the subset of records, belonging to the selected cell which represents a keyword or combination of keywords.

The *Results Analysis Stage* allowed the research team to undertake various combinations of Link Analysis and Dimensional Matrix analyses. The results were then refined based on relevancy (e.g. identifying keywords relating to key knowledge areas of the companies), consistency (e.g. identifying correlation between different keywords relating to key knowledge area with a view to identifying patterns across a number of projects) and completeness (e.g. rules had to be modified to identify a recurring phrase in Company A's reports).

Step 5 consisted of the evaluation and refinement stage. Individual consultations were carried out with both companies to investigate the relevance of the text mining results. This led to a series of actions for both the research team and companies to address in order to improve the quality of the results obtained.

Text Mining Results and Discussion

Three main result sets were obtained for each company. These included (1) Text Analysis, (2) Link Analysis and (3) Dimensional Matrix Analysis. Each of these will be described below.

Text Analysis

The PolyAnalyst software was successful in identifying the frequencies of words from the numerous reports. Table 3 shows an excerpt from Company B.

<Insert Table 3 here>

Text Analysis checks whether the key knowledge areas, previously identified by the experts (see Table 2), can be identified in the reports. For example, generating such analyses would enable companies to look specifically at projects related to these issues, which are

highlighted within the subset of reports e.g. additional costs, budget, etc. This analysis would therefore enable a project team to examine and work together on particular issues which are of most importance to them or the company. The Text Analysis was able to identify specific group of projects that had common issues e.g. changes AND loss and budget AND short lead in times. This will allow companies to examine those PPR reports with common issues in greater detail. One interesting observation is that not all the words identified as *key knowledge areas* by the companies appear in the Text Analysis, or in other words, these topics were not reported at the PPRs.

However, the frequency of occurrence of keywords alone is of limited use and thus the next section goes further to describe the creation and application of rules that could provide potentially more useful results.

Link Analysis

Link Analysis was applied to see if there was a correlation between a set of keywords and phrases under each high level hierarchy identified during the Text Analysis stage. Rules were created using Boolean logic to find topics that may be of interest to the companies. An example of this is “maintain good relationship” AND “profit”. When explored further, the reports indicated that maintaining a good relationship with construction parties lead to a substantial discount and cheap rates which affect the margin or profit on a project. The results of the rule application pilot tests include an abundance of examples of rules that were successful in retrieving specific occurrences of certain keywords and phrases. However, Link Analysis is an area in which domain experience is very important as an expert’s input is needed to identify a set of phrases which should be used to create a useful rule and the expert is subsequently also needed to identify the relevancy of the results and issues of importance. Figure 1 shows the results of the Link Analysis for Company B.

<Insert Figure 1 here>

A large set of correlations, based on the occurrence of keywords and phrases, were found in the reports as shown in Figure 1. Each high level keyword is colour coded and the strength of a correlation is indicated by the thickness of the lines linking the words or phrases. For example, there is a strong correlation between "early stage" and "success", and between "profit" and "value engineering" as well as negative correlations such as "success" and "delay". These correlations could be of interest to the companies and may merit further examination to understand why these terms have such a strong correlation. An example of useful knowledge resulting in Company B's Link Analysis was that out of the 11 project reports examined, five projects faced change of some kind. Of these, three projects incurred loss due to changes; two projects showed that a loss occurred due to design change and expensive redesign at late stages of the project.

However, the major constraints observed during the application of Link Analysis were:

- Repetition of keywords/phrases under each high-level hierarchy; and
- High number of keywords/phrases under each high-level hierarchy.

Therefore, to address these constraints, it was decided:

- To modify the set of keywords/phrases to make each high level hierarchy mutually exclusive i.e. the phrases which appear in one hierarchy will not appear in another.
- Pilot tests should be carried out with the modified and reduced set of key words.
- To apply Link Analysis between one-to-one sets of keywords and phrases from the high level hierarchy. Figure 2 shows the correlation between Time (left) and Finance (right) high level hierarchies for Company A.

<Insert Figure 2 here>

With the approval of the companies, a reduced set of keywords were used to simplify the number of correlations obtained and therefore make the Link Analysis more relevant and easy to understand. The benefits of working with a list of reduced keywords were:

1. Better visual representation of each group during Link Analysis; and
2. Extracting the most useful knowledge based on the listed keywords and phrases as these are most important from companies' viewpoint.

Overall, the Link Analysis results were promising as they showed that some useful (previously unidentified) knowledge can be extracted from the PPR reports. There were useful correlations over several reports that warranted further investigation. Some of the correlations confirm existing knowledge but in other cases, new, previously untapped knowledge was highlighted. Link Analysis correlations is useful for looking at two variables at a time, the next section looks at adding further variables for greater flexibility.

Dimensional Matrix Analysis

A Dimensional Matrix Analysis allows the investigation of a keyword or phrase against many others, for example Finance and Time and Quality. The analysis consisted of setting up columns, each with the heading of a key knowledge area (seven for Company A and four for Company B, each corresponding with a key knowledge area). Dimensional Matrix uses the OLAP - On-Line Analytical Processing feature which provides the capability to perform multi-dimensional analysis of the data. Each column consists of different cells where each cell (block) represents the keyword(s) to be searched for within the PPR reports. While working with the OLAP report, the user defines the values of one or more cells and then browses the subset of records, belonging to the selected cell which represents a keyword or combination of keywords. Another advantage of this approach is that the Dimensional Matrix can be exported to new projects and reused for new datasets.

For example, for Company A, keywords/phrases that come under the Finance column might have occurred with other combination of keywords in different columns (such as Time, Quality, Safety, etc) in the matrix. As shown in Table 4, the phrase “Additional Cost” occurred in six reports. When combined with “Extension of time” in the column Time, both terms appeared in four reports. The knowledge derived can be interpreted as due to the extension of time, four projects incurred additional cost. Furthermore, these two can be combined with Quality to identify how extension of time and additional cost affected quality of product.

<Insert Table 4 here>

In addition to identifying the number of reports, in the lower part of the screen, text mining highlights the occurrences of the keywords within the report to allow the users to investigate further. Thus the Dimensional Matrix Analysis is able to show more complex correlations between multiple key knowledge areas. This would allow organisations to analyse their PPRs based on their multi-dimensional critical success factors.

Evaluation of the Findings

The companies willingly admitted that they had no systems in place to review the findings of their PPR reports and hence any automation of the process would be useful. The results of the Text, Link and Dimensional Matrix Analyses were presented to both companies at a project meeting in order to obtain the initial views on the quality of the results. Both companies found the knowledge provided across a number of projects to be useful and stated that the correlations obtained would warrant further investigation. For example, Company A wished to investigate the correlation between losses on particular type of project that was not previously known. The companies were also very interested in holding further meetings to explore their individual findings in greater detail. Each company requested the research team to refine some of the pilot tests to provide additional results.

Further meetings were arranged to consult each company's representatives separately in order to get a more detailed understanding of whether the results obtained were considered useful and appropriate. The feedback obtained was as follows:

- Companies asked whether the text mining could be directed so that it is possible to see if issues identified get worse or better over time.
- Companies suggested exploring further some vital areas (keywords and phrases) of their company processes through Link Analysis and Dimensional Matrix Analysis to see the outcomes and knowledge to be gained.
- Although many of the results were interesting and promising, some keywords and phrases linked to each other did not appear to be very relevant. The companies suggested refining the list of keywords and phrases to enable a better reflection and focus on relevant areas.
- From the results of the Link Analysis, the fewer number of Company B's reports may have skewed the outcomes and therefore it was suggested that more reports should be provided for further analysis. An additional 10 reports were pre-processed and added to the company's initial results.
- Company A wanted to explore further issues relating to hospital projects as the text mining results suggested that hospital projects consistently incurred a financial loss. This was considered important since the company's strategy was targeted at winning more hospital projects.

Individual consultations were then undertaken within companies. This involved the research team and the company representatives who had identified the key knowledge areas as well as those who conducted the PPR meetings for the companies. This resulted in a list of actions for both the research team and company representatives. Typical actions and outcomes were as follows:

- Revised keywords. Some keywords identified by the Text Analysis were irrelevant and these were therefore deleted or refined. Company B, having reflected on the

results, opted to change their original key knowledge areas. For example, "Communication" was replaced with "Teamwork"; "Communication" then became a sub-set of Teamwork;

- Company B highlighted that the PPR meetings tended to focus on softer issues with a team and customer focus and therefore the hard issues such as Finance and Time did not appear as frequently as expected.
- Thresholds for the frequency of the Link Analysis were increased to show only the links with the highest correlation. For example, Table 5 shows the effect on frequency when a threshold is set for the reports.

<Insert Table 5 here>

The impact of the frequency threshold values is demonstrated in Table 5. If no threshold value is set for a particular keyword e.g. "Quality", all the occurrences of "Quality" across every single report text mined will be reflected in the results of the text mining. This will include both relevant and irrelevant occurrences. When a threshold is set, the text mining process will exclude all occurrences below that threshold. If a keyword appears several times within a report, it is likely that the keyword was relevant to that project. Hence by using the threshold facility, only the reports where the keyword is relevant within a particular context are identified.

Conclusions

A challenge for construction organisations is to obtain an overview of the disparate and numerous review reports stored within the organisation to assist them in providing knowledge that could be used on current or future projects. The experiments investigated whether text mining of Post-Project Review reports could "discover" useful knowledge that can be used for learning of future projects.

A total of 48 reports from two organisations were text mined using Text, Link and Dimensional Matrix Analyses. A number of correlations were obtained across a wide range of reports for all three analyses undertaken. The Link and Dimensional Matrix Analyses were good in identifying issues and correlations occurring across a number of reports. However, there is debate as to whether the correlations identified confirmed existing knowledge or provided "new knowledge". It confirmed existing knowledge in some cases, such as where extensions of time led to a financial loss. However, it also discovered "new" knowledge such as the case where one company was not aware of the repetitive loss on a particular type of project. Discovering this type of "new" knowledge would allow project teams to further investigate trends that occurred in past projects thus making organisational learning from the PPRs less onerous. Another benefit of the text mining approach was it identified which sections of which reports needed to be read when used to assist decision-making. This stimulated discussion within the companies since some of these correlations had never been previously identified and warranted further investigation.

Although text mining proved successful in identifying relationships and some new knowledge; there were limitations to the research undertaken in terms of methodological issues and the quality of the result obtained. The search mechanism was set to identify the *key knowledge areas* that were identified by the companies (see Tables 1 and 2). This therefore restricts the knowledge that can be discovered from the PPRs and new knowledge, not linked to those keywords, yet equally important from a learning perspective, will be missed. The interpretation of the results also needs to be carefully critiqued. Firstly, strong correlations are not automatically important for learning and, vice versa, weak correlations should not automatically be disregarded because these may have a large impact. Secondly, the correlations do not take into consideration the proximity of the keywords. In other words, the correlations show that the words/phrases appear together in the report but they do not indicate whether they appear in the same sentence or paragraph. The final judgements on

the value and usefulness of the identified knowledge must be down to human experience, judgement and in key instances, wisdom.

The implications for future research are that there is a need make the initial text mining stages, including the dictionary of terms used, needs to be easier and more user-friendly. Concurring with the Cross Industry Standard Process for Data Mining (2010) findings, a lot of effort is spent in pre-formatting and converting the text documents and removing unwanted words. Future research should concentrate on how text documents can be submitted in their current format and how the robustness of the correlations can be improved. Also the output, whilst providing useful knowledge, is relatively crude for non-technical users and can be drastically improved to provide more graphical and statistical data on the linkages found. In conclusion, the pilot tests have identified trends, some of which were not previously known, but more work needs to be done before text mining can be widely accepted.

References

Baird, L., Holland, P and Deacon, S. (1999) Learning from action: imbedding more learning into the performance fast enough to make a difference. *Organisational Dynamics,* 27(4), 19-31.

Busby, J (1999) The effectiveness of collective retrospection as a mechanism of organisational learning. *Journal of Applied Behavioural Science,* 35(1), 109-129.

Caldas, C.H. and Soibelman, L. (2003) Automating hierarchical document classification for construction management information systems, *Automation in Construction* 12, pp 395-406.

Caldas, C.H., Soibelman, L. and Han, J. (2002) Automated classification of construction project documents. *Journal of Computing in Civil Engineering* 16(4), pp 234 – 243..

Carrillo, P. (2005) Lessons learned practices in the engineering, procurement and construction sector. *Engineering, Construction and Architectural Management,* 12(3), 236-250.

Carrillo, P.M., Robinson, H.S., Al-Ghassani, A.M. and Anumba, C.J. (2004) Knowledge Management in UK Construction: Strategies, Resources and Barriers", *Project Management Journal*, 35(1), pp 46-56.

Construction Industry Institute (2007) *Effective Management Practices and Technologies for Lessons Learned Programs*, Research Summary 230-1, The University of Texas at Austin.

Cross Industry Standard Process for Data Mining (2010) Process Model available at http://www.crisp-dm.org/Process/index.htm

David Bartholomew Associates (2003) *Learning from experience*, David Bartholomew Associates.

Disterer, G (2002) Management of Project Knowledge and Experiences. *Journal of Knowledge Management,* 5, 512-520.

Easterby-Smith, M. Crossan, M. And Nicoline, D. (2000) organizational Learning: Debates Past, Presnt and Future, *Journal of Management Studies*, 36 (6), 783-796.

Fairclough, J. (2002) *Rethinking Construction Innovation and Research,* Department of Trade and Industry.

Feldman, R. And Dagan, I. (1995) KDT- knowledge discovery in texts. In *Proc. of the First Int. Conf. on Knowledge Discovery (KDD)*, Montreal, Canada, Aug 20-21, pp. 112–117.

Feldman, R. and Sanger, J. (2007) *The Text Mining Handbook: Advanced Approaches to Analyzing Unstructured Data*, Cambridge University Press, Cambridge.

Fellows, R. And Liu, A. (2003), *Research Methods for Construction*, Blackwell Publishing, Oxford.

Gibson, G.E. Caldas, C., Yohe, A. and Weersoriya, R. (2007) *An Analysis of Lessons Learned Practices in the Construction Industry,* Research Report 230-11, Construction Industry Institute.

Harding, J., Shahbaz, M., Srinivas, and Kusiak, A. (2006) Data Mining in Manufacturing: A Review, in American Society of Mechanical Engineers (ASME): Journal of Manufacturing Science and Engineering, 128(4), 969-976.

Hearst, M. (2003), What Is Text Mining? University of California, Berkeley.

Hicks, C.R. (1982) *Fundamental Concepts in Design of Experiments*, Holt-Sanders International, Philadelphia.

Hotho, A., Nürnberger, A. and Gerhard Paaß, G. (2005) *A Brief Survey of Text Mining*, LDV Forum, LDV Forum - GLDV Journal for Computational Linguistics and Language Technology, 20(1),19-62.

Huang, L. and Murphey, Y.L. (2006) Text mining with application to engineering diagnostics. Proc. 19th International Conference on Industrial Engineering and Other Applications of Applied Intelligence (IEA/ AIE 2006), Annecy, France, Jun 27-30. pp 1309-1317.

Kasravi, K. (2004) Improving the engineering processes with text mining. *Proc. of* International Design Engineering Technical Conferences and Computers and Information in Engineering Conference (IDETC/CIE2004). Salt Lake City, Utah, USA Sep 28–Oct 2, pp. 1049-1051

Luhn, H. P. (1958) The automatic creation of literature abstracts, IBM Journal of Research and Development, 2, 159-165.

Nahm, U.Y and Mooney, R. (2002) Text Mining with Information Extraction, American Association for Artificial Intelligence Technical Report SS-02-06.

Prahalad, C. K. and Hamel, G. (1990) The core competence of the corporation. Harvard Business Review, 68(3), 79-92.

Roth, G and Kleiner, A.(1998) Developing organisational memory through learning histories. *Organisational Dynamics,* Autumn, (2), 43-60.

Schindler, M. and Eppler, M.J. (2003) Harvesting project knowledge: a review of project learning methods and success factors. *International Journal of Project Management,* 2, 219-228.

Sowards, D. (2005) The value of post-project reviews. *Contractor,* 52(8), 35-36.

Simon, H. (1969) *The sciences of the artificial*, M.I.T. Press, Cambridge, Mass

Tan, H.C., Carrillo, P.M., Anumba, C.J. and Bouchlaghem, N.M. (2006) Live Capture and Re-use of Project Knowledge in Construction. *Knowledge Management: Research and Practice*, 42, pp 149-171.

Table 1: Company A's list of Key Knowledge Areas

Company A	
Rank	Key Knowledge Area
1	**Financial Issues**
	Additional Costs
	Budget
	Contract Sum
2	**Time**
	Contract programme
	Extension of time
	Lead-in times
3	**Safety**
	Access
	Accidents
	Health & Safety
4	**Environmental Issues**
	Transporting materials
	Waste disposal
	Whole life performance
5	**Quality**
	Aftercare
	Internal audit
	KPI
6	**Trade Packages**
	Drainage
	Drilling
	Electrics
7	**How work was won**
	Competition
	Design and build
	Discounted tender

Table 2: Company B's list of Key Knowledge Areas

Company B	
Rank	Key Knowledge Area
Very High	Change Communication Lead in times Learning Planning
High	Cost Design Information Teamwork Quality
Medium	Budget Delivery Exit margin Negotiation Practical Completion Profit Project completion Scope of work Speed Sub-contractors Tenants
Low	Drainage

Table 3: Result of Text Analysis for Company B

Rule name	Record Count	%
outcome	11	91
learning	10	82
future	9	73
procurement	8	73
response	8	73
stage	8	73
standing	8	73
early stage	8	73
point	8	73
tenant	8	73
use	8	73
end	8	73
works	7	64
way	7	64
access	7	64
attitude	7	64
cause	7	64

Table 4: Dimensional Matrix representing key knowledge areas

Finance	Time	Quality	Health & Safety
Actual cost (15)	Duration (1)	Error (1)	Accident (4)
Additional Cost (6)	**Extension of time (4)**	Success (1)	Protection(2)
Final account (21)	Period (6)	Quality (4)	Fracture (1)
Margin (14)	Stage (6)	Repair (1)	Harm (1)
Price (24)	Tender stage (6)		Injury (2)
Target Cost (4)	Contract period (3)		Reportable accident (1)
Tender Price (7)	Finish (2)		

Table 5: Frequency thresholds: Company B

Keyword	Number of reports without a threshold	Number of reports with threshold limit 3
Quality	20	12
Contractor	20	11
Cost	19	12
Change	18	15
Contract	17	7
Communication	17	10
Completion	16	8
Delivery	15	8
Agreement	15	4
Handover	14	6

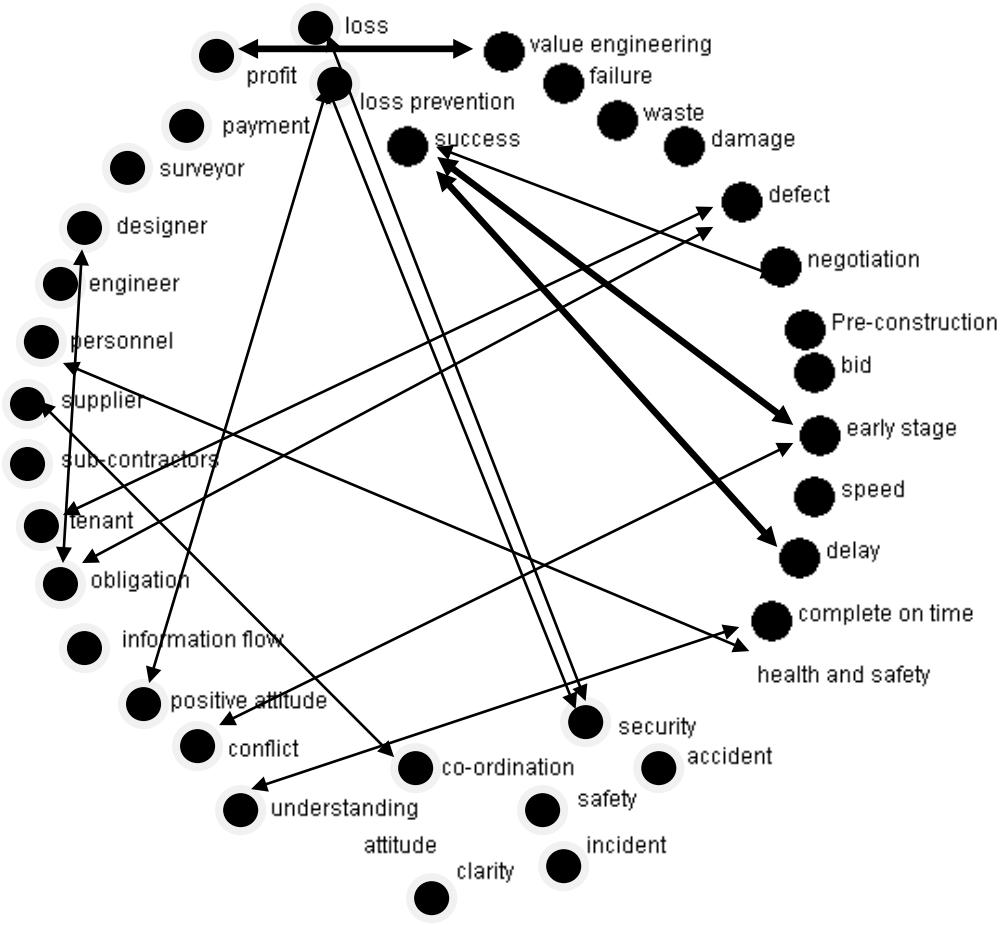

Figure 1: Link Analysis for Company B

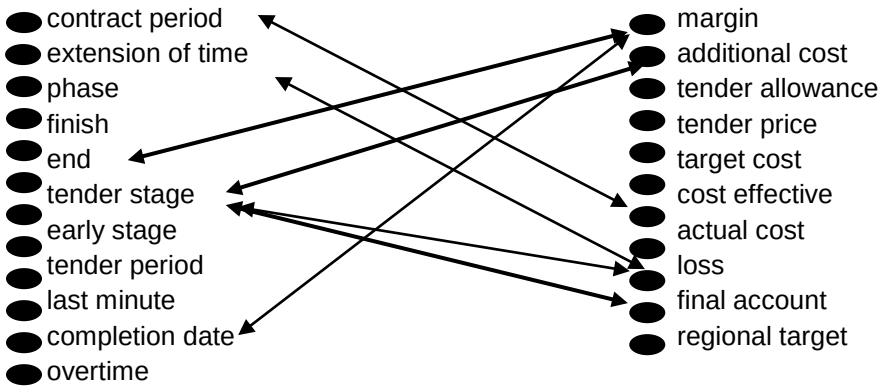

Figure 2: One to One Link Analysis between Time and Finance