Capstone Projects Mining System for Insights and Recommendations

Melvrick GOH, Swapna GOTTIPATI, Venky SHANKARARAMAN
School of Information Systems
Singapore Management University, Singapore
{melvrickgoh.2011, swapnag, venks}@smu.edu.sg

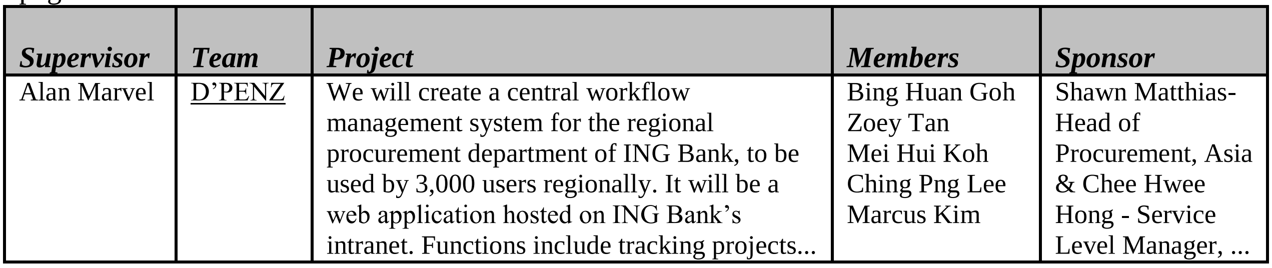
Abstract: In this paper, we present a classification based system to discover knowledge and trends in higher education students' projects. Essentially, the educational capstone projects provide an opportunity for students to apply what they have learned and prepare themselves for industry needs. Therefore mining such projects gives insights of students' experiences as well as industry project requirements and trends. In particular, we mine capstone projects executed by Information Systems students to discover patterns and insights related to people, organization, domain, industry needs and time. We build a capstone projects mining system (CPMS) based on classification models that leverage text mining, natural language processing and data mining techniques.

Keywords: Capstone projects, Data mining, Recommendation system, Visualization, Text analytics, Classification

1. Introduction

An undergraduate capstone project is necessary to help the students cement their knowledge and prepare themselves for real-life work environment. The projects executed by students are usually sponsored by industry professionals who provide requirements, and mentored by faculty who supervise student teams to ensure that the projects are completed on time and satisfy the sponsor requirements. Students usually store project documentations in a common repository. Several existing works propose techniques for assessing students' projects (Rais, et al. 2010), mining the students' projects to discover the students' performance (Bharadwaj, et al. 2011), handling the knowledge of the students' projects (Bender, et al. 2003) and discovering pedagogically relevant knowledge (Merceron et al. 2005). However, very little work has been directed towards mining project documentations which are highly unstructured and textual in nature. Megha, et al. (2014) studied students' projects, where the focus is to track coding repositories to discover insights of students' software development practices.

Mining capstone project documents poses several advantages. Firstly, the project requirements provide an opportunity for discovering current project trends in industry. Knowing industry needs can aid curriculum designers and course instructors to align the course content to emerging trends and therefore equip the students with employability skills. Secondly, when new capstone projects are proposed, the search for similar completed projects can aid students and mentors to plan the techniques, approaches and methods for project execution. Thirdly, for new projects, when there is a need for assigning a mentor for supervision, identifying mentors who have experience in similar past projects will be helpful to ensure success of the project. The main focus of our paper is to mine undergraduate students' capstone project documents to discover insights, and generate summaries and recommendations that aids educators in the decision making process. Figure 1 depicts the sample project information in Wiki pages. We have masked the real names.



Supervisor	*Team*	*Project*	*Members*	*Sponsor*
Alan Marvel	D'PENZ	We will create a central workflow management system for the regional procurement department of ING Bank, to be used by 3,000 users regionally. It will be a web application hosted on ING Bank's intranet. Functions include tracking projects...	Bing Huan Goh Zoey Tan Mei Hui Koh Ching Png Lee Marcus Kim	Shawn Matthias-Head of Procurement, Asia & Chee Hwee Hong - Service Level Manager, ...

Figure 1. A sample project information from capstone projects Wiki page

Supervisor, “Alan Marvel” represents the faculty mentor. Team represents the name of project team and is linked to Wiki page of the project. Project represents a brief description of the project requirements. Comprehensive complete project details are available in the independent Wiki pages. Members represent students in the team. Sponsor, “Shawn Matthias” and “Chee Hwee Hong” represents the project sponsors. Note that the sponsor names are embedded with other details. Now we map the example project from Figure 1 to the different aspects of the mining system. *People* refers to student teams, faculty and project sponsors. *Client* refers to the department and name of the company to which the sponsors belong. The client for this project is “ING Bank N.V.” *Domain* refers to industry to which the company belongs which is not explicit in the Wiki page. For example, “ING Bank N.V.” belongs to the *Banking* domain. *Project needs* or *requirements* represent the requirements given by the sponsors. Project needs are primarily textual in nature and requires a painstaking process to manually analyse full text to discover patterns and insights. Inspired by Sonia, et al. (2010), we propose to generate brief summaries of project needs in terms of keywords or key phrases. Finally, *time* refers to year and semester which are crawled from detailed project Wikis.

One of the major challenges in capstone project documentation is that, it is unstructured and textual in nature. Therefore, we leverage natural language processing and text analytics techniques for processing textual content. In this paper, we describe capstone projects mining system (CPMS), which is a classification based solution. The system takes project data as an input and provides output in five major dimensions; people, client, domain, project needs and time. We built a visualization tool and a recommendation tool as a part of the system. The visualization tool provides interactive graphs that can aid the educators to delve into the details for deeper analysis. Recommendation tool aids in recommending similar projects and project supervisors for new project. We use Information Systems capstone projects (IS480- Information Systems course) to evaluate CPMS. We have collected 322 projects over 8 years span for our system evaluation.

2. Literature Review

Applying data mining techniques in education is an emerging research field. It involves development of methods for making discoveries within the data from educational settings. The goal is to understand students and the learning settings, and to gain insights of educational phenomena (Baker, et al. 2009).

Data mining: Decision trees are used to evaluate students’ performance (Bharadwaj, et al. 2011), for mining learners’ behaviour patterns (Bresfelean, et al. 2007), and tracing deficiencies in students’ understanding (Yoo et al. 2006). Clustering based models are used for detecting the correlation between the students forum participation and final marks of the course (López, et al. 2012), to discover training method for novice learners (Chang et al. 2010), for evaluating students in a tutorial supervisor (Hammouda, et al. 2005), to promote group-based collaborative learning and student diagnosis (Tang et al. 2005). Text mining techniques are used by Gottipati et al (2104) to study Information Systems curriculum for discovering insights in terms of competencies.

Capstone projects: Capstone projects have been studied by various researchers. Many works focus on the design and implementation of capstone courses. Descalu, et al. (2005) studied the project execution procedures in capstone courses. Various tools that assist students and faculty in capstone projects assistance and management were discussed Olarte, et al. (2014). For example, Ceddia, et al. (2002) evaluated project management tool for student projects. Mittal et al. (2014) proposed models based on data mining to discover students’ behaviour by using capstone project coding repositories.

Recommendation systems: Recommendation systems are essential in learning environments for decision making process by students as well as faculty (Hendrik, et al. 2008). Information aggregation using recommendation systems can aid in improving the teaching and learning experiences (Geyer-Schulz, et al. 2001). For example, César, et al. (2009) proposed a collaborative recommendation system based on data mining techniques that can aid students in planning academic itinerary.

Therefore, from the forgoing survey of the related research, it is evident that data mining techniques have been used in education data mining for several tasks related to leaning and evaluation. It is evident that there is need for systems that can mine capstone projects for improving teaching and learning. Inspired by the above works, we mine students’ project repositories for discovering insights from capstone projects. Processing unstructured text requires additional techniques from text mining field for knowledge mining the repository and presenting the results in a meaningful way to the educators.

3. Capstone Projects Mining System

CPMS system is based on data collected from students' capstone Wiki pages which will be inputs to our solution engine. Figure 2 depicts the high level architecture of CPMS system. In next sections, we explain the details of each phase.

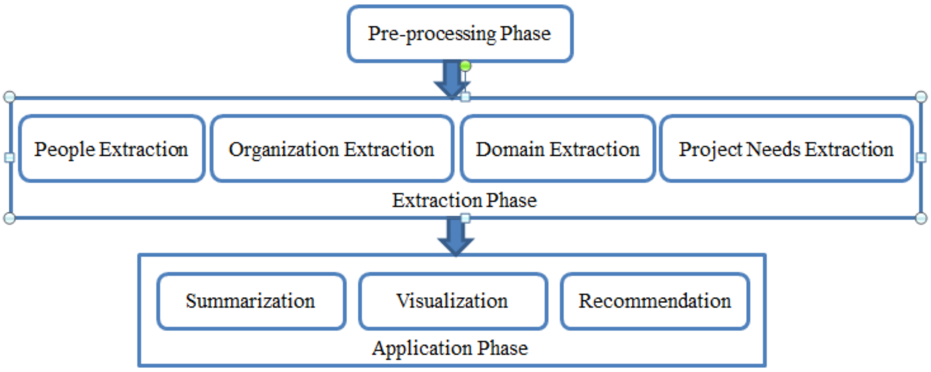


Figure 2. Solution architecture of CPMS

3.1 *Pre-processing Phase*

The objective of pre-processing is to prepare raw text data for the extraction phase. There are several steps: (1) Tokenization involves breaking of a chunk of textual data (e.g., project description) into its constituent words. (2) Stop-word removal removes extremely common words that are of little value in differentiating documents (Salton, et al. 1983). (3) Stemming reduces inflected and derived words into their root word (Porter. 1980).

3.2 *Extraction Phase*

3.2.1 *People and Organization Extraction*

We extract people and organization names by making use of an entity recognition technique. Entity recognition focuses on classifying the different parts of text into different entities (Nadeau et al. 2007) such as names, locations, titles, etc. Figure 3 shows sample client descriptions from our dataset and depict two types of entities; people names (bolded) and organization names (underlined).

(a) **Max XU MengXiang**, Founder/Chief Engineer, GraphPaper Pte Ltd
(b) Avocent Asia Pacific, **Alvin Cheang**, Product Marketing Manager & **Edina Tan**, Field Marketing Manager

Figure 3. Sample client data with people and organization names

3.2.2 *Domain Extraction*

To discover the domain of an organization, we incorporate an external knowledge source, LinkedIn, by virtue of it having one of the most comprehensive records of company data online. In our solution, we exploit LinkedIn's industry API to assign organization names to their industry domains. For example, *Citibank* and *Standard Chartered Bank* are classified under "Banking" domain.

Given an organization name as an input query, LinkedIn API provides a list of matching companies. To discover the best match, we use VSM model (Salton, et al. 1983) to calculate the similarity between organization data from project Wiki and results from LinkedIn. The project data (organization name and project description) as well as the LinkedIn results are conceptually represented by a vector of words. To measure the similarity between these vectors we use cosine similarity. Cosine similarity computes the angle between both vectors which represents their similarity scores (Salton, et al. 1983). In our experiments, we use top ranked result as the matching company and use the industry domain of that company as the domain of the input query.

3.2.3 Project Requirements Extraction

Text mining research provides extraction techniques to extract specific and relevant phrases from unstructured text. Keyphrases are an important means of summarization (Katerina et al. 2000).

Table 1. Project requirements summary of a sample project from our dataset.

Project Description	Project Requirements (Top 7 keyphrases)
The project aims to develop an android-based mobile application for consumers to enjoy targeted discount coupon based on their purchasing …	Android based mobile application, shopping mall, shoppers retail, mobile application, indoor positioning system, facebook integration shoppers…

3.3 Application Phase

3.3.1 Summarization

For summarization, we present the data extracted from extraction phase in a tabular format. Table 2 shows a sample output from our system for people, organization, domain and project needs for selected two projects. The summaries can be further improved using visuals such as word clouds, clusters, etc.

Table 2: Sample structured short summaries by people, organization, domain, and project needs

People	Organization	Domain	Project Requirements
Mishra Nigamananda, Chris Ismael	Standard Chartered Bank iLab	Banking	banking services,snapped qr codes,essential documents,mobile application,fund transfer bill payment,bill splitting,…
Michael Wong, Susan Ngeo, Wing Chew Lau,	Khoo Teck Puat Hospital	Hospital and Health care	health lifestyle analysis system,our system,current manual data entry process,visualization tools,data visualization, data trend....

3.3.2 Visualization

For visualization, we translate the extracted data into a set of interactive graphs that can be used by the school management and educators. The graphs and visuals are based on people, client, domain, project needs and time.

3.3.3 Recommendation

Our recommendation is based on the extraction phase outputs. Given a new project, we first extract people, organization, domain and projects requirements using the extraction phase of the system. We then apply VSM model (Salton et al. 1983) to calculate the similarity scores between the new project and existing projects by comparing their summaries. We then filter existing projects whose similarity scores are above a certain threshold, and sort the projects. The ranked list of similar projects is then recommended by the system for a given new project. Similarly, corresponding supervisors will be recommended by the system.

4. Experiments

We designed our experiments to evaluate our extraction and application phase. Our experiments are designed to answer the following research questions.

RQ1: How effective is extraction phase in extracting people, organization, domain and project needs?

RQ2: How effective is our recommendation phase?

We have in our dataset 322 projects (312 old projects and 10 new projects that start in 2015). In this work, we use Wit.AI[1], an open-source API that provides a classification toolkit, and Apache OpenNLP's[2] POS-tagger. For our experiments, to answer RQ1, we randomly choose around 70% of the projects for training and the rest for testing. To answer RQ2, we recommend relevant old projects and project supervisors to the new projects.

[1] *https://wit.ai/*

[2] *https://**opennlp**.apache.org/*

4.1 Experiment 1: People and organization extraction

We use precision, recall and F-score (Salton et al 1983) for evaluating the effectiveness of our classification-based method for people and organization extraction. These evaluation metrics are computed based on true positives (TP), false positives (FP), and false negatives (FN). We employed a human judge (an undergraduate student majoring in information systems) to identify people and organization names. Table 3 shows the results of our approach. F-score, which is the harmonic mean of precision and recall, is above 96% for people and organization names extraction.

Table 3: People and organization names extraction results

Entity	Precision	Recall	F-Score
People	100.00%	94.05%	96.93%
Organization	100.00%	96.00%	97.96%

4.2 Experiment 2: Domain extraction

We now evaluate our approach on the domain extraction performance. To evaluate the results, we asked a human judge (an undergraduate student majoring in information systems) to label the outputs of our approach as TP, FP and FN. The domain classification results are 76.27% and 56.96% for precision and recall respectively. The F-score of our approach is 65.2% for domain extraction. The F-score for domain extraction is lower than those for people and organization extraction since not all organizations appearing in our dataset exist in LinkedIn. In our analysis, we observed that some are small companies, start-ups, or social enterprises with no LinkedIn pages.

4.3 Experiment 3: Project Requirements Extraction

Recall, that project needs are extracted in the form of keyphrases. We engaged two judges (two SIS undergraduate students) to rate the following statements on a Likert scale of 5 (1 for strongly disagree to 5 for strongly agree):

1. The keyphrases *sufficiently* represent the information contained in the project description.
2. The keyphrases *succinctly* represent the information contained in the project description.

The average rating across both subjects is 4.73 and 4.55 for sufficiency and succinctness respectively. The ratings suggest that keyphrases of project needs are suitable for human consumption.

4.4 Experiment 4: Recommendations

To evaluate our recommendation system, we choose projects for which the system returns at least 5 similar projects. We use average precision to evaluate the results. The average precision is precision averaged across all values of recall between 0 and 1. AP@K is the cutoff at the *k*th result. AP@5 is 71.6% for project manager recommendation. For evaluating the similar projects, we employed a human judge (an SIS undergraduate student) to label each similar project as not relevant (0) or partially relevant (1) or relevant (2). We then aggregate the labels to compute the normalized discount cumulative gain at k (nDCG@k) (Salton et al. 1983) which can handle multiple levels of relevance. Table 6 shows the nDCG@k scores for various values of rankings; *k*. nDCG ensures that highly relevant documents are more valuable than marginally relevant document and vice versa. The result shows that our recommendation system can achieve decent nDCG@1 and nDCG@5 scores of 71.4% and 86.0% respectively.

4.5 Discussions

The five major dimensions namely people, client, domain, project needs and time dimensions described in Section 3 for extraction phase are not exhaustive and can be extended to others such as product names, technologies, time zones, languages, etc. We ignore the positions and titles of the people in this study. For future work, extending into other facets of classification is useful for deeper insights and for generating better summaries of the projects. LinkedIn API provides additional attributes of a company such as location, employee size, etc. The additional attributes offer visibility over the more granular aspects of the software projects for decision making process. We leave this as an interesting future work. We acknowledge that this current model of matching to LinkedIn data is still open to plenty of improvement and we propose the following measures for improving the accuracy.

5. Conclusion

This paper describes an educational projects mining system (CPMS). The main goal of CPMS is to mine student project repositories to discover insights and present the insights in visuals that are easy to interpret by the educators or school management. The insights from textual content of project documentations are extracted by using textual analytics, natural language processing and classification techniques. We are currently exploring topic models for extracting keyphrases for project needs. A big gap in our approach is that images and videos were not in the current work. Going forward we would like to take into consideration these types of data as part of our mining efforts.

References

Baker, R.S.J.d., Yacef, K. (2009). The State of Educational Data Mining in 2009: *A Review and Future Visions. Journal of Educational Data Mining*, 1 (1), 3-17.

Bender B. & Longmuss J. (2003). Knowledge Management in Problem-Based Educational Engineering Design Projects, *In International Journal of Engineering Education. (19) 5. Dublin, pp. 706-711.*

Bharadwaj B.K. and Pal S. (2011). Mining Educational Data to Analyze Students' Performance, *International Journal of Advance Computer Science and Applications (IJACSA)*, Vol. 2, No. 6, pp.63-69.

Bresfelean, V.P. ; Babes-Bolyai Univ., Claj-Napoca. 2007. Analysis and Predictions on Students' Behavior Using Decision Trees in Weka Environment. *In ITI 2007.* 51-56

Ceddia, J. and Sheard, J. (2002). Evaluation of WIER - A capstone project management tool. International Conference on Computers in Education (ICCE 2002), Auckland, New Zealand.

César Vialardi Sacín, Javier Bravo Agapito, Leila Shafti, Alvaro Ortigosa.2009. Recommendation in Higher Education Using Data Mining Techniques. In EDM 2009, Cordoba, Spain, July 1-3, 2009.

Chang, Lin and Bai, Xue (2010). Data Mining: A Clustering Application. In proceedings of 14th Pacific Asia Conference on Information Systems, PACIS. Paper 184.

Christopher D. Manning. (2011). Part-of-speech tagging from 97% to 100%: is it time for some linguistics? In Proceedings of the (CICLing'11), Alexander Gelbukh (Ed.), Vol. Part I. , Berlin, Heidelberg, 171-189.

Descalu, S., Varol, Y., Harris, F. Westphal, B. (2005). Computer Science Capstone Course Senior Projects: From Project Ideato Prototype Implementation. In FIE'05. Proceedings 35th Annual Conference(pp. S3J-1).

Geyer-Schulz, M. Hahsler and M. Jahn, Educational and Scientific Recommender Systems: Designing the Information Channels of the Virtual University, Int. J. Eng. Educ. 17(2), 2001, pp. 153±163.

Hammouda, K., Kamel, M. (2005) Data Mining in e-Learning. *In Samuel Pierre (Ed.),* Springer-Verlag, Berlin Heidelberg New York (2005).

Hendrik Drachsler, Hans G. K. Hummel, and Rob Koper. 2008. Personal recommender systems for learners in lifelong learning networks; the requirements, techniques and model. Int. J. Learn. Technol. 3, 4 404-423.

Katerina Frantzi, Sophia Ananiadou, Hideki Mima (2000). Automatic Recognition of Multi-Word Terms: the C-value/NC-value Method. International Journal of Digital Libraries, 3(2) pp.117-132.

López, M. I., Luna, J. M., Romero, C., Ventura, S. (2012). Classification via clustering for predicting final marks starting from the student participation in forums. In EDM (pp.148–151)

Megha Mittal and Ashish Sureka. 2014. Process mining software repositories from student projects in an undergraduate software engineering course. In (ICSE Companion 2014). ACM, New York, USA, 344-353.

Merceron, A., Yacef, K. (2005). Educational Data Mining: a Case Study. *In International Conference on Artificial Intelligence in Education*, Amsterdam, The Netherlands, 1-8.

Nadeau D., Sekine S. (2007) A survey of named entity recognition and classification. Lingvisticae Investigationes, 30:3–26.

Nigam, K.; Lafferty, J. & McCallum, A. (1999), Using maximum entropy for text classification, in IJCAI-99.

Olarte, J.J., Dominguez, C., Jaime, A., Garcia-Izquierdo, F.J.(2014). A tool for capstone project management in computer science engineering. International Symposium on Computers in Education (SIIE)

Porter M.F. (1980). An algorithm for suffix stripping, Program, 14 (3):130-137.

Rais, S.S.; Enzai, N.; Kar, S.A.C.; Aris, M.A.; Kadir, E.A.; Darus, R. (2010) Assessments of final year diploma electrical engineering project 2. *Engineering Education (ICEED), 2nd International Congress*, 139 – 144

Salton, G. and McGill, M.J. (1983) Introduction to modern information retrieval. McGraw Hill, New York.

Sonia Haiduc, Jairo Aponte, Laura Moreno, and Andrian Marcus. (2010). On the Use of Automated Text Summarization Techniques for Summarizing Source Code. *In WCRE*. Washington, DC, USA, 35-44.

Gottipati. S, Shankararaman. V. (2014). Analyzing Course Competencies: What can Competencies Reveal about the Curriculum?. In 22nd International Conference on Computers in Education. (ICCE'14).

Yoo, J., Yoo, S., Lance, C., Hankins, J. (2006). Student Progress Monitoring Tool Using Treeview. In proceedings of the 37th Technical Symposium on Computer Science Education, SIGCSE'06. ACM Press. 373-377