Mining Of Text In The Product Development Process

Han Tong Loh, Rakesh MENON, Christopher K Leong

***Abstract*-- In the prevailing world economy, competition is keen and firms need to have an edge over their competitors for profitability and sometimes, even for the survival of the business itself. One way to help achieve this is the capability for rapid product development on a continual basis. However, this rapidity must be accomplished without compromising vital information and feedback that are necessary. The compromise in such information and feedback at the expense of speed may result in counter-productive outcomes, thereby offsetting or even negating whatever profits that could have been derived. New ways, tools and techniques must be found to deliver such information. The widespread availability of databases within the Product Development Process (PDP) facilitates the use of data mining as one of the tools. Thus far, most of the studies on data mining within PDP have emphasised on numerical databases. Studies focusing on textual databases in this context have been relatively few. The research direction is to study real-life cases where textual databases can be mined to obtain valuable information for PDP. One suitable candidate identified for this is "voice of the customer" databases.**

I. Product Development Process (PDP)

A good product development process is a key to creating a successful product. A well- organized and coherent development process serves to ensure the efficient delivery of a final product that suits customer's wants. Such products are truly the lifeblood of a company's long term economic existence. Thus it is no surprise that companies are willing to invest both time and effort to ensure a proper product development process so as to deliver competitive products.

A Product Development Process is the sequence of steps or activities which an enterprise employs to conceive, design and commercialize a product (Ulrich and Eppinger, 2000). Although every organization may follow a slightly different process, the basic elements are usually the same. In essence, the major steps that would usually be incorporated into the PDP are:

Han Tong Loh, is with the Innovation in Manufacturing Systems and Technology(IMST), Singapore-MIT Alliance (SMA), N2-B2c-15, Nanyang Technological University, Nanyang Avenue, Singapore 639798

Rakesh Menon, is with Centre for Robust Design, Department of Mechanical Engineering, National University of Singapore,10 Kent Ridge Crescent, Singapore 119260

Christopher K.Leong, is with the Innovation in Manufacturing Systems and Technology(IMST), Singapore-MIT Alliance (SMA), N2-B2c-15, Nanyang Technological University, Nanyang Avenue, Singapore 639798

- Planning
- Design
- Production
- Service and Support

At each step different milestones have to be met. It is almost inevitable that several problems would be faced before these milestones could be reached or sometimes modified (due to the inability to achieve them). A wide variety of tools are currently used in industry to address some of these problems. Syan (1994) presented such a list of tools in his paper for a seven-phase PDP. These are shown in Table 1, with some slight modifications.

Table 1: Tools used within the PDP

Tools	PD step
QFD Problem-Solving Techniques Design for Manufacture	Planning
QFD Product FMEA Taguchi Design for Manufacture	Design
QFD Problem-Solving Techniques Process FMEA Taguchi Process Capability Modelling SPC	Production
Problem-Solving Techniques	Service and Support

These tools have been found to be very effective in solving problems faced in the PDP. However, with increasing competition and challenge, companies are seeking the use of new technologies and methodologies to improve their PDP. This fact, coupled with the recent explosion of information technology that has enabled companies to collect and store increasing amounts of information, has given rise to the use for a collection of new techniques, popularly known as data mining.

The usefulness of such techniques is so well recognized that studies calling for re-engineering of business processes and incorporation of data warehousing to facilitate data mining has emerged (Mulvenna et. al, 1996; Cheng and Chang, 1998). These techniques perform best when massive amounts of data are available. Under these circumstances, manual processing of such data becomes inefficient and, in many cases, even impossible.

II. Background on Data Mining

A. What is Data Mining?

Data mining, as defined by Fayyad et al (1996a), is the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data. This is a revised definition from that of Frawley et al (1991) to reflect developments and growth in data mining. According to the latter, it is the non-trivial extraction of implicit, previously unknown, and potentially useful information from data. Many people treat data mining as a synonym for another popularly used term, knowledge discovery in databases(KDD). On the other hand, others view data mining as simply an essential step in the KDD process.

Knowledge discovery, which is an iterative process, is depicted in Figure 1 (Han and Kamber, 2000). According to this view, data mining is only one step in the entire process, albeit an essential one, since it uncovers hidden patterns for evaluation. However, in industry as well as in the database research milieu, these two terms are often used interchangeably. A detailed description of the steps in the data mining process is given by Fayyad et al (1996b).

B. Data Mining Operations

Depending on the objective(s) of an analysis, different types of data mining operations could be used. In general, these data mining operations are used for characterizing the general properties of a database or for performing inference on the data in order to make predictions. (Cabena et al, 1997; Berry and Linoff, 1997). Some of the more commonly used operations are outlined in Table 2.

Table 2: Common operations in data mining

Operation	Description
Predictive Modeling	Predicts a continuous or discrete output, given a particular input
Clustering	Automatically clusters homogenous groups of data
Association Analysis	Find the links between items in transactions
Deviation Detection	Detects outliers

C. Applications

There have been numerous applications of data mining in a variety of disciplines ranging from agriculture to finance. Table 3 from Frawley et al (1992) is not intended as an exhaustive list but has been included to demonstrate the range and diversity of applications. Besides many others, Shortland and Scarfe (1994) have added to this list by describing how British Telecoms had used data-mining techniques for various applications within the information and communications industry.

Table 3: Applications of Data Mining Techniques

Discipline	Application Examples
Medicine	Biomedicine, drug side effects, hospital cost containment, genetic sequence analysis and prediction
Finance	Credit approval, bankruptcy, stock market prediction, securities, fraud detection, detection of unauthorized access to credit data, mutual fund selection
Agriculture	Soya bean and tomato disease classification
Social	Demographic data, voting trends, election results
Marketing and sales	Identification of socio-economic subgroups showing unusual behaviour, retail, shopping patterns, product analysis, frequent flying patterns, sales prediction
Engineering	Automative diagnostic expert systems, computer aided design databases, job estimates
Physics and chemistry	Electrochemistry, superconductivity research
Law	Tax and welfare fraud, fingerprint matching, recovery of stolen cars

The overwhelming number of applications of data mining in various disciplines is highly apparent. However, for the sake of this study, focus would be on the literature related to the application of data mining within the product development process.

III. Data Mining and the PDP

A. Previous Work on Data Mining within the PDP

Alex et al (1997) outlined steps to be undertaken for implementing data mining within a manufacturing environment. They briefly mentioned the use of data mining for fault diagnosis, process and quality control, and machine maintenance tasks. They also mentioned that, in the near future, data mining has the potential of becoming one of the key components in manufacturing scenarios.

Dagli and Hsi (1997) studied the impacts of data mining technology on product design and process planning, emphasizing current practices in manufacturing. They mentioned that the final product design impacts 60 to 80 % of the total manufacturing costs. Their emphasis was on the use of the World Wide Web and associated protocol, to create an environment for exchange of data and information, at real time, in a wide scale.

Another application of data mining for the failure diagnosis of process units using probabilistic networks and decision trees was reported by Wang et. al (same comments)(1997). A semi-automatic approach was used. A commercial software - C5.0 - was used to predict a case class from its attribute values. In this paper a failure database was investigated.

Koonce et al (1997) presented a software, DBMine, for searching databases for learning from manufacturing systems. The tool implemented three common data mining techniques : Bacon's algorithm, Decision Trees and DB-Learn. The authors used DBMine to investigate job-shop schedules produced by Genetic Algorithms (GA) to determine patterns in the gene sequencing, as well as rules that help explain these patterns.

The nonlinear prediction of manufacturing systems, using data mining techniques, was investigated by Kim and Lee (1997). They examined the relationship between the inputs and outputs using techniques such as regression, moving averages, exponential smoothing, neural networks and case-based reasoning. They also examined the robustness of the learning techniques under varying patterns of noise.

Ferguson et al (1998) claimed that the provision of information relevant to various functions of the enterprise can be achieved by the introduction of an effective decision support mechanism. The primary goal of the paper was to establish data mining as a mechanism which enables information to be used, from later life-cycle stages, by earlier ones, as well as to provide this information in a format which would be understandable and useable to another product life-cycle function. Therefore they assessed the technique of data mining as a means of providing decision support. They identified two databases to which data mining methods could be applied. One database consisted of coded customer feedback reports from beta-tests and failed products already in use by customers. The other database contained the material features on every product object.

In an application of fault detection, Milne et al (1998) used historical data for early detection of paper defects so that corrective action could be applied. They used decision trees for on-line prediction of the quality of the reels of paper.

The use of data mining in a chemical industry was investigated by Mastrangelo and Porter (1998). They used classification and regression trees (CART) for selecting key process variables as inputs to a multivariate statistical process monitoring (SPM) system. The objective was to reduce the number of breakages of the mono-filament fibres which they were producing.

In an attempt to improve process yield, McDonald (1999) applied data mining tools for yield improvement in integrated circuit manufacturing. He used data mining to study the correlation between process variables and yield, to detect defects caused by the sequence in which the wafers were processed, and to relate spatial arrangement of defects on the wafer to fault causes.

Unlike most of the previous works that dealt primarily with quantitative data, the study by Tan et al (2000), investigated service center call records comprising both textual and fixed-format columns, to extract information about the expected cost of different kinds of service requests. They found that the incorporation of information from free-text fields provided a better categorization of these records, thus facilitating better predictions of the cost of the service calls.

As can be seen from the surveyed literature thus far, many of the studies carried out have concentrated on quantitative databases within the PDP. In particular, most of the applications seem to be centered on manufacturing processes. Little work has been done to explore the potential of textual databases within the PDP, although there exists an immense amount of such data.

B. Textual Data within the PDP

Figure 2 displays typical information found within the PDP. This information was gathered from studying two Multi-National Corporations (MNCs) in Singapore. It could be deemed to be representative of data usually collected in the PDP for a consumer product.

As can be seen from this figure, a huge amount of data is available. For example, in the production stage, as much as 30,000 units could be produced each week. Another noteworthy point is the fact that different types of databases are available within the process. For example, in the planning and designing stage, the data is available in a report form. Both pictures and sentences are present. Such databases are sometimes referred to as multi media databases. In the production and service stages, both numerical and textual databases can be found.

As such, textual data bases are available at various stages of the PDP, and it would be a loss for firms not to capitalize on the benefits to be derived from the mining of such databases.

C. The Need for Mining Textual Databases within PDP

Current methods of analyzing textual data within the PDP includes the use of spreadsheets and manual processing, to decipher meaning and relationships from the textual input. Such tasks usually entail a lot of time and resources, which could be otherwise better utilised. Hence it would be necessary, if not extremely useful, to have automated or semi-automated text analysis schemes that would be able to infer important and pertinent information out of such huge and intimidating databases.

Quantitative databases are relatively easy to mine and there are already various established techniques for this. In comparison, textual databases/fields are much more difficult to manipulate and there is a greater level of difficulty in mining such databases. Hence a lot of textual databases within the PDP end up simply as archives. However, a vast amount of valuable information and knowledge lies dormant, only awaiting some tool such as textual data mining, in order to tap into their potential.

One might argue that the encoding of texts could be

employed to deal with textual databases, hence avoiding textual data mining. However, many problems exist with respect to this. Firstly, such encoding can take a long time. Understanding the different possible problems that have occurred with the product, classifying such problems, and finally encoding them, is no easy task. Secondly, with rapid innovation in today's industries, products in the market change very rapidly. Every time a design-change in the product occurs, the encoding system needs to be modified or, in some instances, even changed completely. It is actually possible that the product in question might have finished its market-life before such changes in the encoding system have been incorporated. Thirdly, even if an encoding list is available, it might be too long that personnel using it tend to bypass using the encoding list for some other quick alternatives (This disturbing trend has been observed for one of the databases investigated in this study). Finally, although it is possible for free-texts to contain a lot of unnecessary content and remarks, one could still obtain certain significant information details from them that a structured and rigid encoding system would not facilitate. As such, encoding systems can never serve as a perfectly good substitute for free-texts. However, a hybrid of both may prove to be useful.

IV. Conclusion

Textual data are found at different stages of the PDP. For the time being, much of the information and knowledge that can be extracted from such databases lie dormant, waiting to be tapped. Hence there is great potential for such information to be used effectively. One potential use is for a better understanding of the PDP as well as for product improvement. However, such analysis of textual data is not frequently done within the manufacturing/engineering discipline. This may largely be due to a lack of awareness of tools that are usually more popular and commonly found in the Computer Science community. This is an unfortunate state. Besides drawing attention to this current state of affairs, this paper has highlighted the tools and techniques in data mining and summarized briefly the few pieces of published work that have been done in the area of data mining in the product development process.

References

[1] Berry. M.J. and Linoff, G. *Data Mining Techniques ; For Marketing, Sales and Customer Support*. First ed.: John Wiley and Sons, 1997.

[2] Buchner, A.G., Anand, S.S. and Hughes, J.G. "Data Mining in Mining in Manufacturing Environments: Goals Techniques and Applications", *Studies in Informatics and Control* 6,no.4 (1997): 319-328

[3] Cabena, P., Hadjinian, P., Stadler, R., Verhees, J. and Zanasi, A. *Data Mining from Concept to Implementation*: Prentice Hall, 1997.

[4] Cheng, P.S. and Chang, P. "Transforming Corporate Information into Value through Data Warehousing and Data Mining." *Aslib Proceedings* 50, no. 5 (1998): 109-113.

[5] Dagli, C.H. and Lee, H-C, ed. *Impacts of Data Mining Technology on Product Design and Planning*. Edited by F. Plonka and G.Olling, Computer Applications in Production and Engineering: Chapman & Hall, 1997, pp.58-70.

[6] Fayyad, U.M., Piatetsky-Shapiro, G. and Smyth, P, ed. *From Data Mining to Knowledge Discovery*. Edited by G. Piatetsky-Shapiro U.M. Fayyad, P. Smyth and R. Uthurusamy, Advances in Knowledge Discovery and Data Mining: AAAI/MIT Press, 1996a, pp.1-33.

[7] Fayyad, U.M., "Data Mining and Knowledge Discovery: Making Sense out of Data" *IEEE Expert* 1(5) (1996b) 20-25.

[8] Ferguson, C-J., Lees, B., MacArthur, E. and Irgens, C. "An Application of Data Mining for Product Design." A paper delivered at the IEE Colloquium on Knowledge Discovery and Data Mining, 1998, pp.5/1-5/5

[9] Frawley, W.J., Piatetsky-Shapiro, G., Matheus, C.J. "Knowledge Discovery in Databases; An Overview." *AI Magazine* Fall 1992, pp.57-70.

[10] Frawley, W.J., Piatetsky-Shapiro G. and Matheus, C.J., ed. *Knowledge Discovery in Databases: An Overview*. Edited by G. Piatetsky-Shapiro and B.Frawley, Knowledge Disovery in Databases: AAAI/MIT Press, 1991.

[11] Joachims, T. *Text Categorization with Support Vector Machines : Learning with Many Relevant Features*. : University of Dortmund, 1998, LS-8 Report 23.

[12] Han, J. and Kamber, M. *Data Mining; Concepts and Techniques*: Morgan Kaufman, 2000.

[13] Keerthi, S.S, Shevade, S.K., Bhattacharyya, C, and Murthy, K.R.K. *A Fast Iterative Nearest Point Algorithm for Support Vector Machine Classifier Design* : Indian Institute of Science, 1999, TR-ISL-99-03.

[14] Kim, S.H. and Lee, C.M. "Nonlinear Prediction of Manufacturing Systems through Explicit and Implicit Data Mining." *Computers and Industrial Engineering* 33, no. 3-4 (1997).

[15] Koonce, D.A., Fang, C-H. and Tsai, S-C. "A Data Mining Tool for Learning from Manufacturing Systems." *Computers and Industrial Engineering* 33, no. 1-2 (1997): 27-30.

[16] Lovins, J. "Development of a Stemming Algorithm." *Mechanical Translation and Computational Linguistics* 11 (1968): 22-31.

[17] Mastrangelo, C. M., Porter, J. M. "Data Mining in a chemical process application; *Proceedings of the 1998 IEEE International Conference on Man, Systems and Cybenatics* 3 (1999): pp2917-2922.

[18] Mcdonald C.J. "New tools for yield improvement in integrated circuit manufacturing: can they be applied to reliability? *Microelectronics Reliability*39 (1999) 731-739.

[19] Milne, R., Drummond, M., Renous, P. "Predicting paper making defects on-line using data mining"; *Knowledge-Base System* 11 (1998): 331-338

[20] Mulvenna, M.D., Büchner, A.G., Hughes, J.G. "Re-engineering Business Processes to Facilitate Data Mining." A paper delivered at the Proceedings of 1st International Conference on Practical Aspects of Knowledge Management, Basel, Switzerland, 1996.

[21] Porter, M. "An Algorithm for Suffix Stripping." *Program* 14, no. 3 (1980): 130-137.

[22] Salton, G. and McGill, M.J. *Introduction to Modern Information Retrieval*. First ed.: McGraw-Hill International Book Company, 1983.

[23] Salton, G. and Buckley, C. "Term Weighting Approaches in Automatic Text Retrieval." *Information Processing and Mangement* 24, no. 5 (1988): 513-523.

[24] Shortland, R. and Scarfe, R. "Data Mining Applications in BT." *BT Technological Journal* 12, no. 4 (1994): 17-22.

[25] Syan, C.S. "Introduction to Concurrent Engineering." In *Concurrent Engineering; Concepts, Implementation and Practice*, ed. C.S. and Menon Syan, U.: Chapman & Hall, 1994.

[26] Tan, P-N., Blau, H., Harp,S. and Goldman, R. "Textual Data Mining of Service Centre Call Records." A paper delivered at the The Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Boston, MA, USA, 2000.

[27] Ulrich, K.T. and Eppinger, S.D. *Product Design and Development*. Second Ed., 2000.

[28] Van Rijsbergen, C.J. *Information Retrieval*: Butter Worths, 1979.

[29] Wang, X.Z., Chen, B.H. and McGreavy, C. "Data Mining for Failure Diagnosis of Process Units by Learning Probabilistic Network." *Transactions of Institute of Chemical Engineers* 75, no. Part B (1997): 210-216.

Figure 1: Knowledge Discovery Process