INTERNATIONAL CONFERENCE ON ENGINEERING DESIGN, ICED'07
28 - 31 AUGUST 2007, CITE DES SCIENCES ET DE L'INDUSTRIE, PARIS, FRANCE

KNOWLEDGE DISCOVERY AND MANAGEMENT FOR PRODUCT DESIGN THROUGH TEXT MINING – A CASE STUDY OF ONLINE INFORMATION INTEGRATION FOR DESIGNERS

Ying Liu[1], Wen Feng Lu[2] and Han Tong Loh[2]

[1] The Hong Kong Polytechnic University
[2] National University of Singapore

ABSTRACT

Product innovation and design is often regarded as an information and knowledge intensive activity. While existing tools, i.e. search engines and product data management systems, have provided a certain degree of assistance to design community in information storage, processing and retrieval, many challenges remain, especially if we intend to provide advanced capabilities so that the designers are able to handle the overloaded information. These capabilities include, but are not restricted to, automated classification of design relevant documents based on product architecture, relevance analysis and summarization of design documents, identification of product function from customer feedback\reviews in conceptual design, personal design information subscription and management, market analysis and competitive intelligence report generation and so on. To fulfill these demands, an integrated text mining system for knowledge discovery and management in product design is proposed. In this paper, we report our study on the integration of online information with the internal knowledge base, e.g. product taxonomy, as part of our efforts towards the realization of the proposed system. Several key techniques are explained, i.e. maximal frequent word sequence algorithm in discovering the quality phrases, document profile model based on salient semantic information and finally concept based automated text classification approach. Experimental studies have demonstrated its effectiveness.

Keywords: Knowledge Discovery and Management, Product Design, Text Mining, Text Classification, Online Information

1 INTRODUCTION

Intensive global competition is pushing manufacturing companies ever harder in their strife for constant profit. As the world economy evolves from a capital based paradigm to a knowledge centric one, manufacturing companies are increasingly concerned about the acquisition, management and utilization of advanced R&D information and knowledge from both internal organizations and external resources, e.g. design documents and customer feedbacks or e-journals and digital libraries. For one example, product design engineers are interested in design experience, technical tips and solutions of earlier product models, which may have been written down in documents by previous engineers. For another example, the customer manager from service department intends to check whether customers' complaints are related to hardware failures or their improper operations. Such complaints are manually logged in text forms by service operators. Successfully handling of such textual information can enrich the company's understanding about the current market, save the development cost, shorten the time-to-market, bring better products to satisfy customers, and in turn lead the company to prosperity [1-4].
Within the last decade, Text Mining (TM) has emerged as a very exciting research area as well as a promising technology to "discover by computer of new, previously unknown information, by automatically extracting information from different written resources" [5]. While it remains arguable what exactly TM is, it can be simply considered as the extended application of data mining over textual data, such as emails, web pages, and other digital documents. Researchers who share this view actually define TM more from the perspective of tasks associated, such as text classification and text

clustering. As a matter of fact, TM is a fast evolving area which is supported by several knowledge branches in computer science and artificial intelligence, e.g. information retrieval (IR), machine learning (ML), natural language processing (NLP), and database technology. Due to the very nature of human languages, TM is inherently more difficult than data mining where the latter only focuses on numerical data. Problems like word ambiguity, different representations of similar concepts in human natural language, and the implicit semantic information resident in textual data will remain as challenges as well as opportunities for many years to come [5-9].

This paper serves for two purposes. In the first place, we would like to introduce a knowledge discovery and management (KDM) system for product design based on text mining methodology. Secondly, as part of our continuous efforts to research and implement the proposed KDM system, we report our study on online information integration for designers through automated text classification. We first briefly review the related work and practice of KDM in engineering design in Section 2. After the introduction of KDM framework in Section 3, we present a case study. Several key techniques are explained in detail in Section 4. Experimental results are reported in Section 5. Section 6 concludes.

2 RELATED WORK

Since the early 1990's, Knowledge Management (KM) has gained worldwide recognition of creating an organization's core competency [10, 11]. With such a transformation from a material and energy founded industrial world towards a knowledge based economy [12], it has become a grand challenge to effectively acquire, process, share, and manage the individual and corporate knowledge in nurturing innovative ideas.

In the context of engineering, design is the core topic to KM research and practice. Basically, at least two streams of approaches are well known in facilitating KM in design, i.e.

- From the organization or individual's behaviour point of view [13-18]
- From the information and communication technology perspective, such as product data management and collaborative design environment [19-25]

As highlighted by McMahon [26] that currently several crucial challenges exist, these include, but are not limited to, a large amount of undigitized design information in textual format, the urgent demand for automated text classification of design related texts, the in-depth study of design documents which often has to call upon advanced linguistic techniques and the overloaded information on designers.

In our research, we emphasize on the technology aspects of KM in design. We introduce the latest research findings in data mining and TM, also known as knowledge discovery from unstructured data, and tailor them according to the specific requirements of design environment. Different from the existing PDM systems and collaborative design environments, the TM based approach stresses the potential of discovery and management functions, e.g. search interesting and meaningful patterns, extract internal links (e.g. social network) based on the analysis of linguistic clues hidden in the design documents, and classify documents based on the predefined thematic categories. The whole approach has initiated a new paradigm for KM research and practice in engineering design. However, by saying so, we still acknowledge the very importance of behaviours based study and exercises. As a matter of fact, both approaches often interact with each other. For instance, one designer's experience with TM based tools will surely affect the way he\she requests, analyzes and shares information and knowledge with his\her colleagues.

3 KDM FOR PRODUCT DESIGN THROUGH TEXT MINING

Throughout the product design process, various information and knowledge are utilized, e.g. strategic information, market and customer feedback, technical tips and know-how, competitors, and regulatory [3, 27]. Such information is often presented in textual format, e.g. memos, emails and web pages, and is stored in different locations, from the internally developed and owned enterprise database to those external ones.

In view of the challenges imposed by the large amount of textual data and the distributed location of information and knowledge sources needed, we propose an integrated text mining system for product design. Several core tasks which offer great assistance to design personnel are introduced. Figure 1 shows the framework of our KDM system which provides a comprehensive support for industrial firms to undertake the KDM initiatives.

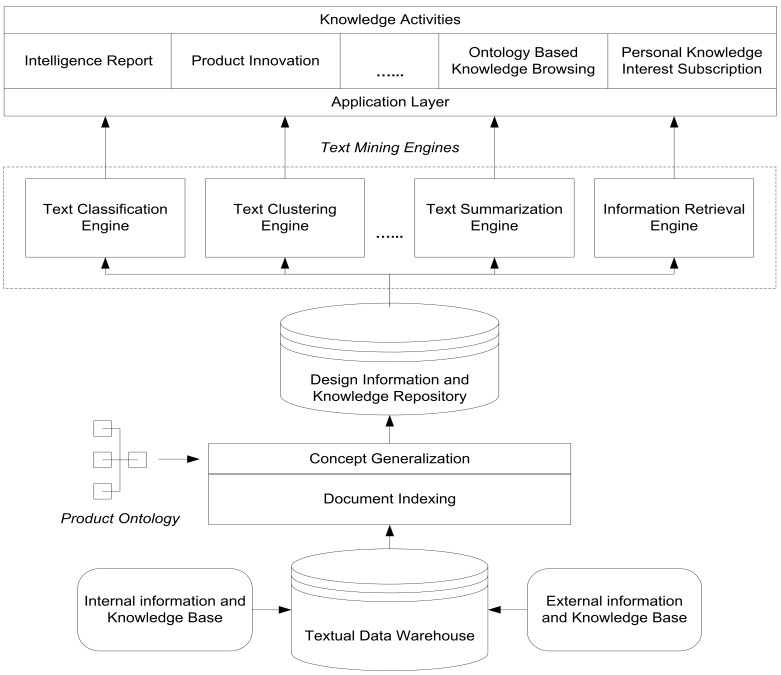

Figure 1. The System Architecture of Knowledge Management for Product Design through Text Mining

The system integrates the information and knowledge from both internal and external sources first. Different from most existing information retrieval (IR) systems, such as search engines, the layer of concept generation and representation is built based on the enterprise knowledge taxonomy and provides the critical function to represent enterprise knowledge concepts in an explicit way. Therefore, the collected documents are distilled and indexed in a concept-based space based on these knowledge concepts and finally stored in the knowledge repository. While the system does offer the classic IR function, which is more for ad-hoc information searching purpose, various text mining engines are created to support the company-wide knowledge activities. Built upon the knowledge repository, mining engines refer to a series of subsystems which each fulfills a dedicated major function, e.g. document classification based on knowledge taxonomy, single or multiple documents summarization, document clustering and many more. Finally, various knowledge activities are carried out based on the structured knowledge repository and the support of knowledge mining engines. These include, but are not limited to, competitive intelligence report generation, product innovation and technical tip sharing, ontology based knowledge browsing, personal knowledge interest subscription, etc. These mining engines, operating in a concept-based space, are considered as the core components providing the critical role for information and knowledge management purpose according to designers' needs.

4 ONLINE INFORMATION INTEGRATION FOR DESIGNERS

Online information is one sort of external information sources that is often deemed as indispensible to product designers in gathering relevant intelligence and aspiration. For instance, online customer reviews with respect to various commercial products contain the most valuable information which directly describes customers' experience using the products, their preferences and opinions. Product designers often gain insights from their manual analysis of such comments, and therefore, are able to understand their customers better and improve their products accordingly. However, the rapid advancement of WWW, information technology and e-commerce has made the Internet forums, e-opinions portals and personal blogs widely accessible to consumers. As a result, it has overwhelmingly changed the ways how customers express their concerns as well as how designers should gather such information. While the Web has become a rich resource of consumer opinions, the sheer amount of data and their distributed locations have emerged as a grand challenge facing product designers.

In this section, we record our study on the integration of online information with the internal knowledge base, e.g. product taxonomy, as part of our research efforts towards the realization of the aforementioned KDM system. Figure 2 visualizes the basic idea of our approach. As noted, designers are searching information from different online sources. These include, for example, customer reviews

found in Amazon (www.amazon.com) and Epinions (www.epinions.com), latest products announced in the competitors' corporate websites, relevant breakthroughs publicized at the information portals of research institutes and so on. Our system is expected to capture the salient semantic information from various pieces of textual information. Finally, a concept based text classification system will judge the relevance of different information retrieved from online sources and then categorize them into the product taxonomy accordingly.

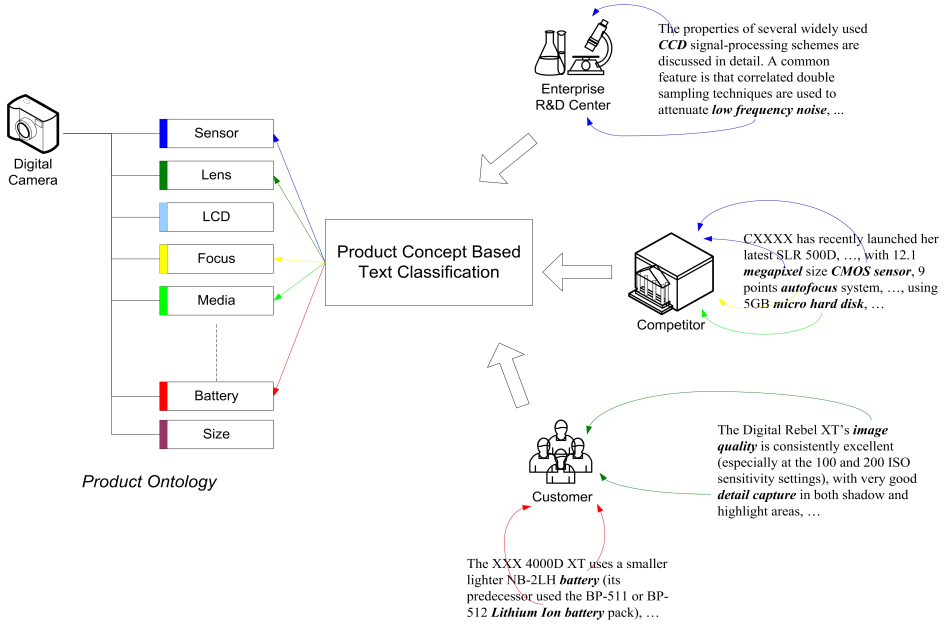

Figure 2. Integration of Online Information for Designers

In the following subsections, we focus on the explanation of several key initiatives undertaken. These include:

- Maximal frequent word sequence: an algorithm that is able to generate a quality set of phrases which represents the core semantic information of documents.
- Document profile model: a semantic based document modelling, built upon maximal frequent word sequences, which is able to represent documents effectively and highlight the semantic relation among documents.
- Concept based text classification: an attempt which aims to explore the potentials of automated text classification based on the document profile model.

As for the issue of information retrieval from various websites, a comprehensive strategy has been developed, e.g. manually identify relevant targets, consult Yahoo directory, and filter the results from Google etc. We omit the detailed description in this paper largely because of its nature in system development and integration and the page limit.

4.1 Maximal Frequent Word Sequence

The idea and method of discovering the set of maximal frequent word sequences (MFSs) out of a textual dataset was first proposed by Ahonen [28]. A MFS is a sequence of words that is "frequent in the document collection and, moreover, that is not contained in any other longer frequent sequence" [28, 29]. A word sequence is frequent if it appears in at least σ documents, where σ is a pre-specified support threshold. The goal of the MFS algorithm is to find all maximal frequent phrases in the textual dataset. The strength of the method is that it employs a versatile technique for finding sequential text phrases from full text, allowing, if desired, gaps between the words in a phrase [30]. For example, the word sequence "*product knowledge databases*" can be extracted as a frequent phrase even if its occurrence is in the form of:

- "…*product* management using *knowledge databases*…"
- "…*product* data in *knowledge databases*…"
- "…*product* specifications, *knowledge databases*…"

in the supporting documents. The maximum gap allowed between words of a sequence is determined by the maximal word gap parameter. Algorithm 1 explains the basic logic of MFS.

Algorithm 1: Discovery of maximal frequent sequences in the documents

```
Input:     S: a set of pre-processed documents, σ: a support threshold,
           g: maximal word gap
Output:    Max: a set of maximal frequent sequences

// Initial phase: collect all frequent pairs
1.For all the documents d ∈ S
2. Collect all the ordered pairs and occurrence information within d
3.Grams2 = all ordered pairs that are frequent in S
  // Discovery phase: build longer word sequences by expanding and joining and store MFSs
    and occurrence information into Max
4.k := 2;
5.Max := ∅;
6.While Gramsk is not empty
7.  For all grams g ∈ Gramsk
8.      If g is not a subsequence of some m ∈ Max
9.          If g is frequent
10.                 max := Expand(g);
11.                 Max := Max ∪ max;
12.                 If max = g
13.                         Remove g from Gramsk;
14.         Else
15.                 Remove g from Gramsk;
16.     Prune(Gramsk);
17.     Gramsk+1 := Join(Gramsk);
18.     k := k + 1;
19. Return Max
```

The algorithm starts with a set of ordered frequent word pairs, such as *ab* and *bc*. Then it searches for the MFSs in a bottom-up and greedy manner. It adds an item into a pair in a greedy manner by trying all possible choice, until the longer sequence is no longer frequent. As this process repeats, it only expands the sequences that are not contained in any of the existing maximal sequences. If some cannot be expanded, they are themselves maximal sequences. It stops until there are no grams left for expansion.

To the best of authors' knowledge, the generation of MFSs is conceptually similar to the well known Apriori algorithm [31-33]. Although in Ahonen's initial work [28], no connection between MFS and Apriori has been mentioned, MFS is indeed dedicated to mine sequential textual patterns. However, different from Apriori, we do not cap the maximal length of word sequences and we need to specify the word-gap-in-between in MFS. In the original proposal and its application of Apriori, e.g. basket data mining, the gap-in-between may not play a significant role in identifying the shopping pattern of a customer. For example, the manager may not care about when or in what exact sequence that the customer picks up beers and snacks. However, we are afraid that's not true in word sequences. A loose pattern, i.e. a word sequence with a wide gap among its words, will reduce the effectiveness of using MFSs to represent the core content of documents. In fact, it has generated more meaningless sequences.

4.2 Document Profile: Semantic Based Document Modelling

Borrow from Information Retrieval, the most widely accepted document model in both text classification and clustering is probably vector space model [34-36], i.e. a document d_i is represented as a vector of term weights $\vec{v}_i(w_{1i}, w_{2i}, ..., w_{|\Gamma|i})$, where Γ is the collection of terms that occur at least once in the document collection D. This is often named as bag-of-words (BoW). The term weights $w_{|\Gamma|i}$ is often computed in *tfidf* manner, i.e. *term frequency* times *inverse documents frequency*, [37-40].

Besides the vector space model, some other commonly used models are Boolean model, probability model [41, 42], inference network model [43], and statistical language model [34]. Essentially, a statistical language model is concerned about the probabilities of word sequences, denoted as $P(S : w_1, w_2, ..., w_n)$. These sequences can be phrases, clauses and sentences. Their probabilities are often estimated from a large text corpus. As reported, statistical language modelling has been successfully applied to many domains, such as its original application in speech recognition and part-of-speech tagging [44], information retrieval [45-47] and spoken language understanding [48]. In practice, the most widely used statistical language model is n-gram model [34]. The n-gram model is formed based

on the observation that the occurrence of a specific word may be affected by its immediately preceding words. In unigram model, words are assumed to be independent. Hence, the probability of a word sequence *S* is approximated by the product of all individual words' probabilities. In bigram and trigram model, more context information has been taken into consideration, i.e. the probability of word depends on its previous one word or two words respectively.

However, the limitation with BoW approach is apparent. Single words are not the only units to convey the thematic information. While we observe that sequences like "finite element analysis", "machine learning" and "supply chain management" are surely in a better position to represent the content of sentences and documents, they are rarely integrated into the current study and application in text processing, automated text clustering and classification [34, 35, 49].

In order to resolve this difficulty, we propose a document modelling approach called Document Profile (DP) [50]. In essence, DP is concerned about how documents should be represented. In the BoW approach, documents are basically defined by all single words which have occurred at least once in the document collection. In our work, DPs are given by a set of fine selected single words Γ and *MFS* discovered. The particular questions are how to search them and what should be included in documents' DPs.

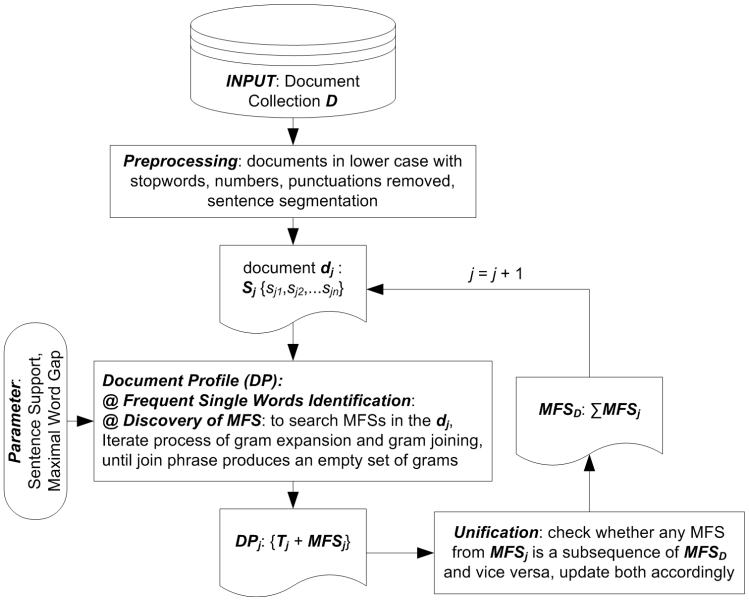

Figure 3. The General Flowchart of Document Profile Discovery [50]

Figure 3 illustrates how we generate the document DPs and index documents from the system point of view. Suppose we have generated $DP_{j+1}\{MFS_{j+1}+\Gamma_{j+1}\}$ for document d_{j+1}, the unification step will examine whether any MFSs in MFS_j is a subsequence of any MFSs in MFS_{j+1} and vice versa. For instance, "wireless network" and "wireless network mobile computing" are two MFSs from MFS_j and MFS_{j+1} respectively. Apparently, "wireless network" is a subsequence of the latter, and hence, it will be replaced by the latter in MFS_j with its occurrence information being left intact. As a result, the system dynamically maintains an overall MFS list, MFS_D, which covers every single MFS found in the document collection *D*. This examination and unification step is deemed as the key to provide more semantic clues with regard to the document's thematic information and retain the dimension of term space small.

4.3 Concept Based Text Classification

The typical approach of automated text classification (TC) is often dominated by supervised learning based techniques [34, 51]. It is mainly involved with two phases, i.e. document indexing and classifier induction [35]. In the last decade, automated TC has achieved very promising results. Many machine learning based techniques have been applied to TC tasks, including Support Vector Machines (SVM) [52, 53], *k*-Nearest Neighbors (*k*NN) [54], decision trees and decision rules [55, 56], and Bayesian related techniques [57, 58] and so on. Its application has been extended to documents organization, text filtering, word sense disambiguation, etc [35].

However, existing TC tasks mainly consider single terms as the means to convey the information. Since document representations have been largely confined to term frequencies, these approaches ignore the rich information resident in the semantic relations among terms. As pointed out by Dumais, "interest rate" might be a better index rather than either "interest" or "rate" [59]. Surprisingly, previous efforts in generating more sophisticated document representation have not shown significant improvement than those bag-of-words based approaches [35]. These include the work to use n-grams, phrases, or co-occurred words as context information to enrich the representation [58-60], or to tackle the linguistic variation such as synonyms, vocabulary and word choice by introducing techniques like synonym clusters and Latent Semantic Analysis (LSA) [61, 62]. We noted that some of these approaches are either expensive to compute or the results are not human readable.

From the point of view of knowledge management, we favor the work which extends the current term based representations with n-grams, phrases, and co-occurred words which are all in a human readable form. Due to this reason, we have examined the previous work [58, 63] using phrases as the index terms. We repeated their procedures to generate the phrases in several datasets. The major problems we noted are many phrases are actually irrelevant to the themes of categories and the term space has often become extremely large afterwards. Furthermore, our TC experiments using phrases alone as the indexes cannot compete with those only using single terms. This alerts us that the phrases should be carefully selected rather than a simple combination of words and more attention should be paid to integrate such phrases in a more sensible way as well.

While there are many different ways to express and handle concepts, we take a simple approach by using the concept-centered phrases, i.e. MFSs, plus all single terms as the descriptors to represent and manipulate the knowledge concepts in automated text processing tasks [64]. In our study, we apply the state-of-the-art TC algorithms over the proposed DP model. Since DP model has offered a concise document representation which possesses salient semantic information, we intend to find out whether the DP model is effective in chasing better performance in TC.

5 EXPERIMENTAL STUDY

We carry out the experimental study to demonstrate how the integration of online information with internal knowledge can be achieved through text classification based on our proposed approach, i.e. the DP model plus concept based TC, and the prevailing approach in automated TC, i.e. the BoW model. Manufacturing Corpus Versions 1 (MCV1) was utilized to represent the internal knowledge base [65]. MCV1 is an archive of 1434 English language manufacturing related engineering papers which we gathered by the courtesy of the Society of Manufacturing Engineers (SME). It combines all engineering technical papers published by SME from year 1998 to year 2000. All documents were manually classified. Figure 4 visualized the concept hierarchy and the category distribution of MCV1 which were adopted in our study.

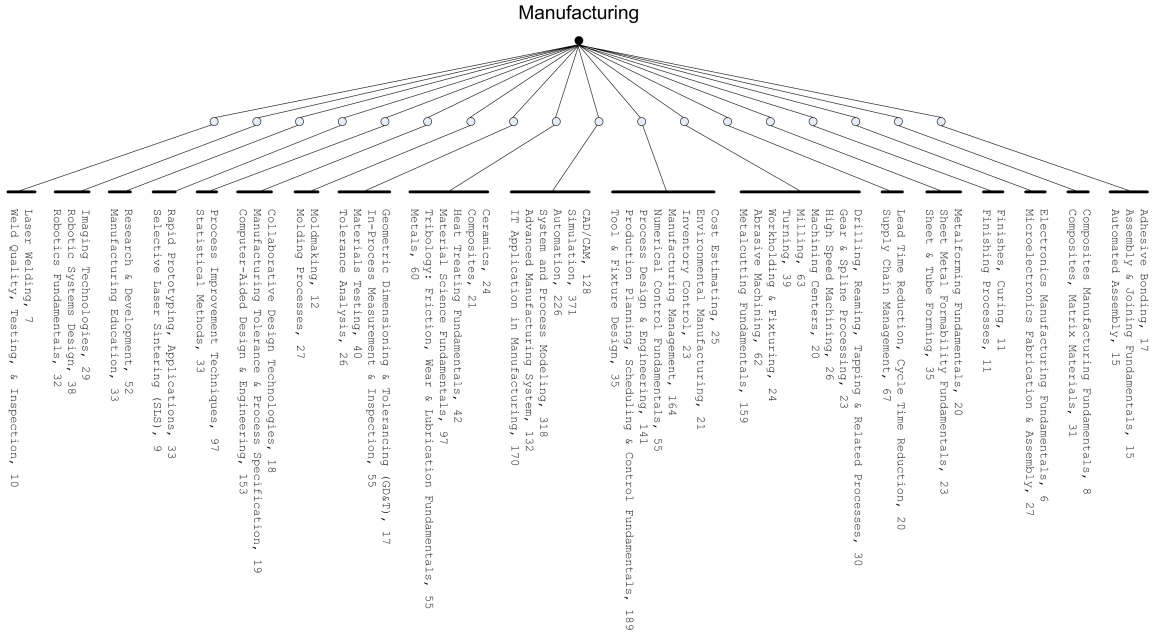

Figure 4. The Concept Hierarchy of MCV1 Tested in the Experiment

Besides MCV1, a collection of 500 abstracts of technical papers were retrieved from various relevant websites, e.g. Society of Manufacturing Engineers (SME), American Society of Mechanical Engineers (ASME) and Institute of Industrial Engineers (IIE). This collection serves as the online information to be integrated with MCV1 and its concept hierarchy. Therefore, the text classifiers were trained and tested over MCV1, whereby these 500 abstracts are considered as unseen data for integration.
The state-of-the-art Support Vector Machine (SVM) [53] was chosen as the classification algorithm, in particular, the well known implementation SVM^{Light} [52]. Linear function has been adopted as its kernel function, since previous work has shown that the linear function can deliver even better performance without tedious parameter tuning in TC [52, 66]. Major standard text preprocessing steps were applied in our experiments, including tokenization, stop word and punctuation removal, and stemming. However, feature selection was skipped for SVM experiments and all terms left after stop word and punctuation removal and stemming were kept as features.

Table 1. A Contingency Table for Evaluation in a Binary Setting, where 'a' is the Number of Items that Match a Method's Outputs and Target Values, i.e. the Number of Outputs Correctly Categorized, 'b' and 'c' are the Number of Items that Mismatch a Method's Outputs and Target Values Respectively

	Expert Yes	Expert No
Output Yes	*a*	*b*
Output No	*c*	*d*

As for the performance measurement, a contingency table for a topic category, as shown in Table 1, is adopted to measure the results by means of the following two standard metrics, i.e. the F_1 values and false alarm [39, 40, 67] which are defined as: recall: $r=a/(a+c)$, precision: $p=a/(a+b)$, F_1: $F_1=2pr/(p+r)$ and false alarm: $f=b/(b+d)$. Owing to the reality that the 500 abstracts retrieved only belong to 17 categories among those 63s shown in Figure 4 (*ground truth*), we decide to include false alarm as an additional metric in evaluating the system performance. The idea is for the cases where documents are actually from that specific category, F_1 should be reported. A higher F_1 value simply means that more documents are correctly categorized. On the contrary, false alarm is implemented to examine the 46 categories left, meaning all documents are supposed to be classified as *No* with respect to any of these 46s. A lower false alarm value indicates the fact that more documents have been correctly rejected.

Assembly and Joining	*Robotics and Machine Vision*
assembl line assembl process assembl sequenc assembl system genet algorithm line balanc ……	ccd camera imag process machin vision parallel manipul robot system six degre freedom … …
Forming and Fabricating	*Rapid Prototyping*
bend oper die cast draw die draw process finit element method form sheet metal heat transfer laser cut process metal flow pressur die process model process paramet process plan roll form sheet metal form process trial error ……	comput aid design layer manufactur rapid prototyp technolog revers engin select laser sinter sl solid freeform stereolithographi apparatu sla ……
	Welding
	arc weld ga metal laser weld weld paramet weld process ……

Figure 5. Examples of MFS Discovered

Figure 5 shows some MFS examples generated from various categories in MCV1. These MFSs are very close to the quality of human defined domain thesaurus. We have observed that incorporating more semantic information in DP models has a significant impact on the similarity computation. For instance, in BoW approach documents centered on "Forming & Fabricating", "Assembly & Joining"

and “Welding” are usually dominated by a few frequent words, such as “process”, “metal” and “manufacturing” and so on. These words do not offer an effective means to represent the difference among documents. In contrast, the integration of MFSs, e.g. “draw process”, “assembly process”, “laser cut process”, “weld process”, “sheet metal form process” and many others, have brought more clues to link up the similar ones.

Figure 6 reports the overall performance of text classification, in terms of F_1, that integrates the online information, i.e. 500 abstracts, and the internal knowledge, i.e. MCV1, based on the prevailing BoW approach and our DP modeling approach. Apparently, our proposed DP model plus concept based text classification have demonstrated its superior performance compared to the classic BoW model across all 17 categories, regardless of the quantity of documents available. A close look on the document representation reveals that using DP model has effectively reduce the dimension of term space from a figure of slightly more than seven thousand to less than two thousand. Meanwhile, many salient semantic clues appear to be available. This leads us to believe that DP model has played a key role to improve the results.

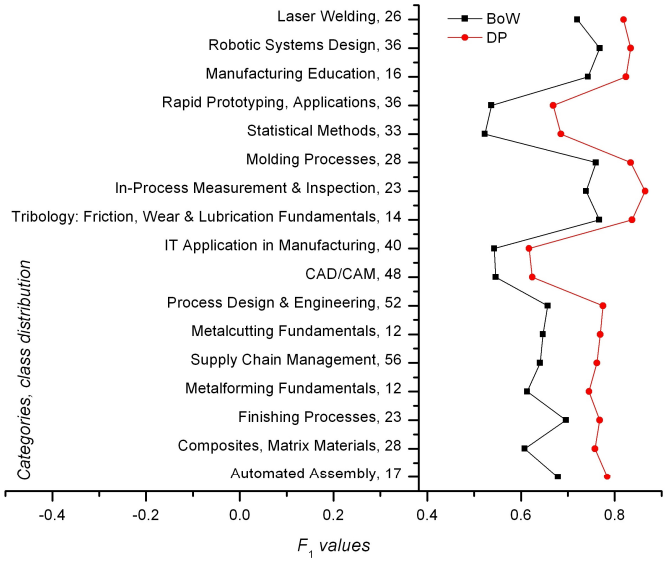

Figure 6. Overall Classification Performance, i.e. F_1, Based on BoW Approach and DP modeling Approach using SVM

The assessment of other 46 categories in terms of false alarm further convinces us the effectiveness of DP modelling approach. The ideal situation is that every of these 46 classifiers should not accept a single document from the set of 500 abstracts. We note that when DP model was adopted the averaged false alarm has dropped to slightly higher than 5%, in contrast to BoW’s 11% in our study. This implies that using DP model can exclude a significant amount of erroneous information.

6 CONCLUSION

In this paper, a framework of knowledge discovery and management for product design through text mining has been introduced. We provide a case study on the integration of online information with the internal knowledge base as part of our efforts in researching and implementing the proposed KDM system in design environment. Several key initiatives undertaken are described in detail. These include maximal frequent word sequence algorithm that is introduced to generate a quality set of phrases which are able to represent the core semantic information of documents; document profile model - a semantic based document modelling built upon MFSs which is able to represent documents effectively and highlight the semantic relation among documents, and finally concept based text classification - an attempt to explore the potential of automated text classification based on document profile model. The experimental study has shown the effectiveness of our approach in integrating external information through classification and its capability of excluding erroneous information. This has encouraged us to further explore other text based mining tasks using document profile model.

REFERENCES

[1] Liu, Y., A Concept-based Text Classification System for Manufacturing Information Retrieval. *Singapore MIT Alliance*, 2005, (National University of Singapore)

[2] Braha, D., ed. *Data Mining for Design and Manufacturing: Methods and Applications*, 2001 (Kluwer Academic Publishers

[3] Liu, Y., Lu, W.F. and Loh, H.T., A Framework of Information and Knowledge Management for Product Design and Development – A Text Mining Approach. *Proceedings of the 12th IFAC Symposium on Information Control Problems in Manufacturing, INCOM'06*, 2006,

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

[5] Hearst, M.A., Untangling Text Data Mining. *Proceedings of ACL'99, the 37th Annual Meeting of the Association for Computational Linguistics, invited paper*, 1999,

[6] Dixon, M., An Overview of Document Mining Technology, 1997, (Computer Based Learning Unit, University of Leeds)

[7] Merkl, D., Text Data Mining, In Dale, R., Moisl, H. and Somers, H., eds. *A Handbook of Natural Language Processing - Techniques and Applications for the Processing of Language as Text*, 1998, (Marcel Dekker)

[8] Rajman, M. and BESANÇON, R., Text mining: natural language techniques and text mining applications. *Proceedings of the seventh IFIP 2.6 Working Conference on Database Semantics (DS-7)*, 1998,

[9] Visa, A., Technology of Text Mining, In Petra Perner, ed. *Proceedings of Machine Learning and Data Mining in Pattern Recognition, Second International Workshop, MLDM 2001*, 2001, pp.1-11 (Springer)

[10] Stewart, T.A., *Intellectual Capital: The New Wealth of Organisations*, 1997 (Doubleday)

[11] Edvinsson, L. and Malone, M.S., *Intellectual Capital: Realizing Your Company's True Value by Finding Its Hidden Brainpower*, 1997 (Collins)

[12] Drucker, P., *The Age of Discontinuity: Guidelines to Our changing Society*, 1969 (Harper and Row)

[13] COURT, A.W., CULLEY, S.J. and MCMAHON, C.A., Information access diagrams: a technique for analysing the usage of design information, *Journal of Engineering Design*, 1996, 7, 55-75

[14] MARSH, J.R., The capture and structure of design experience. *Department of Engineering*, 1997, (University of Cambridge)

[15] BROOKES, N.J., SMART, P. and LETTICE, F., The importance of informal networks to effective design management. *Design Management, Process and Information Issues, Proceedings of the International Conference of Engineering Design* 2001, pp.433-440 (Professional Engineering Publishings)

[16] BADKE-SCHAUB, P. and FRANKENBERGER, E., Analysis of design projects, *Design Studies*, 1999, 20, 465-480

[17] BADKE-SCHAUB, P. and FRANKENBERGER, E., Design representations in critical situations of product development. *Proceedings of the 4th Design Thinking, Design Representation Symposium*, 1999,

[18] CHAN, W.K. and MAUBORGNE, R., Fair process: managing in the knowledge economy, *Harvard Business Review*, 2003, 127-136

[19] ELLIS, C.A., GIBBS, S.J. and REIN, G.L., Groupware: some issues and experiences, *Communications of the ACM*, 1991, 34, 39-58

[20] GOMES, S., EYNARD, B. and MAGNON, L., A multi-site design experience using a computer supported collaborative work tool. *Proceedings of the International Conference on Engineering Design, ICED01*, 2001, pp.379-386 (Professional Engineering Publishing)

[21] MCGREGOR, S., THOMSON, A.I. and JUSTER, N.P., Information sharing within a distributed, collaborative design process: a case study. *Proceedings of the ASME Design Engineering Technical Conferences*, 2001, (American Society of Mechanical Engineers)

[22] LIU, T.D. and XU, W.X., A review of web-based product data management systems, *Computers in Industry*, 2001, 44, 251-262

[23] Lowe, A., McMahon, C. and Culley, S., Characterising the requirements of engineering information systems, *International Journal of Information Management*, 2004, 24(5), 401-422
[24] LOWE, A., MCMAHON, C.A., SHAH, T. and CULLEY, S.J., A method for the study of information use profiles for design engineers. *Proceedings of the ASME Design Engineering Technical Conference & Computers and Information in Engineering Conference*, 1999, pp.DTM-8753 (American Society of Mechanical Engineers)
[25] HAAS, R.E., Engineering knowledge management-current status and future challenges. *Proceedings of the 6th International Conference on Concurrent Enterprising*, 2000, pp.161-164
[26] Mcmahon, C., Lowe, A. and Culley, S., Knowledge management in engineering design, personalisation and codification, *Journal of Engineering Design*, 2004, 15(4), 307-325
[27] Zahay, D.L., Griffin, A. and Fredericks, E., Exploring Information Use in Detail in the New Product Development Process. *Proceedings of PDMA research forum*, 2003, pp.55-72
[28] Ahonen-Myka, H., Finding All Frequent Maximal Sequences in Text. *Proceedings of the 16th International Conference on Machine Learning ICML-99 Workshop on Machine Learning in Text Data Analysis*, 1999, pp.11-17
[29] Yap, I., Loh, H.T., Shen, L. and Liu, Y., Topic Detection Using MFSs. *Proceedings of the 19th International Conference on Industrial & Engineering Applications of Artificial Intelligence & Expert Systems (IEA\AIE 2006)*, 2006,
[30] Ahonen-Myka, H., Heinonen, O., Klemettinen, M. and Verkamo, A.I., Finding Co-occurring Text Phrases by Combining Sequence and Frequent Set Discovery. *Proceedings of 16th International Joint Conference on Artificial Intelligence IJCAI-99 Workshop on Text Mining: Foundations, Techniques and Applications*, 1999,
[31] Agrawal, R., Imieliński, T. and Swami, A., Mining association rules between sets of items in large databases. *Proceedings of the 1993 ACM SIGMOD international conference on Management of data*, 1993, pp.207-216
[32] Agrawal, R. and Srikant, R., Fast Algorithms for Mining Association Rules, In Bocca, J.B., Jarke, M. and Zaniolo, C., eds. *Proceedings of the 20th Very Large Data Bases (VLDB) Conference*, 1994, pp.487-499 (Morgan Kaufmann)
[33] Agrawal, R. and Srikant, R., Mining Sequential Patterns. *Proceedings of the Eleventh International Conference on Data Engineering*, 1995, pp.3-14
[34] Manning, C.D. and Schütze, H., *Foundations of Statistical Natural Language Processing*, 1999 (The MIT Press)
[35] Sebastiani, F., Machine Learning in Automated Text Categorization, *ACM Computing Surveys (CSUR)*, 2002, 34(1), 1-47
[36] Jurafsky, D. and Martin, J.H., *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*, 2000 (Prentice Hall)
[37] Salton, G. and Buckley, C., Term Weighting Approaches in Automatic Text Retrieval, *Information Processing and Management*, 1988, 24(5), 513-523
[38] Salton, G. and McGill, M.J., *Introduction to Modern Information Retrieval*, 1983 (McGraw-Hill)
[39] Baeza-Yates, R. and Ribeiro-Neto, B., *Modern information retrieval*, 1999 (Addison-Wesley Longman Publishing Co., Inc.)
[40] van_Rijsbergen, C.J., *Information Retrieval*, 1979 (Butterworths)
[41] Robertson, S.E., The Probability Ranking Principle in IR, *Journal of documentation*, 1977, 33(4), 294-304
[42] Fuhr, N., A Probabilistic Model of Dictionary Based Automatic Indexing. *Proceedings of the riao 85 (Recherche d' Informations Assistee par Ordinateur)*, 1985, pp.207-216
[43] Turtle, H. and Croft, W.B., Inference networks for document retrieval. *Proceedings of the 13th annual international ACM SIGIR conference on Research and development in information retrieval*, 1989, pp.1-24
[44] Charniak, E., *Statistical Language Learning*, 1993 (MIT Press)
[45] Miller, D.R.H., Leek, T. and Schwartz, R.M., A Hidden Markov Model Information Retrieval System. *Proceedings of SIGIR-99, 22nd ACM International Conference on Research and Development in Information Retrieval*, 1999, pp.214-221

[46] Ponte, J.M. and Croft, W.B., A Language Modeling Approach to Information Retrieval. *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1998, pp.275-281
[47] Hiemstra, D., A Linguistically Motivated Probabilistic Model of Information Retrieval. *Research and Advanced Technology for Digital Libraries, Second European Conference, ECDL '98*, 1998, pp.569-584
[48] Zue, V.W., Navigating the Information Superhighway Using Spoken Language Interfaces, *IEEE Expert: Intelligent Systems and Their Applications*, 1995, 10(5), 39-43
[49] Jain, A.K., Murty, M.N. and Flynn, P.J., Data clustering: a review, *ACM Computing Surveys (CSUR)*, 1999, 31(3), 264-323
[50] Liu, Y., Loh, H.T. and Lu, W.F., Deriving Taxonomy from Documents at Sentence Level, In Prado, H.A.d. and Ferneda, E., eds. *Emerging Technologies of Text Mining: Techniques and Applications*, 2007, (Idea Group Inc.)
[51] Mitchell, T.M., *Machine Learning*, 1997 (The McGraw-Hill Companies, Inc.)
[52] Joachims, T., Text categorization with Support Vector Machines: Learning with many relevant features. *Machine Learning: ECML-98, Tenth European Conference on Machine Learning*, 1998, pp.137-142 (Springer)
[53] Vapnik, V.N., *The Nature of Statistical Learning Theory*, 1999 (Springer-Verlag)
[54] Yang, Y. and Liu, X., A re-examination of text categorization methods. *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, 1999, pp.42-49 (ACM Press New York, NY, USA)
[55] Apté, C., Damerau, F. and Weiss, S.M., Text Mining with Decision Trees and Decision Rules. *Conference on Automated Learning and Discovery*, 1998,
[56] Quinlan, J.R., *C4.5: Programs for Machine Learning*, 1993 (Morgan Kaufmann)
[57] Friedman, N., Geiger, D. and Goldszmidt, M., Bayesian network classifiers, *Machine Learning*, 1997, 29(2-3), 131-163
[58] Lewis, D.D., An evaluation of phrasal and clustered representations on a text categorization task. *Proceedings of SIGIR-92, 15th ACMInternational Conference on Research and Development in Information Retrieval*, 1992, pp.37-50
[59] Dumais, S., Platt, J., Heckerman, D. and Sahami, M., Inductive learning algorithms and representations for text categorization. *Proceedings of the seventh international conference on Information and knowledge management*, 1998, pp.148-155 (ACM Press New York, NY, USA)
[60] Sahlgren, M. and Cöster, R., Using Bag-of-Concepts to Improve the Performance of Support Vector Machines in Text Categorization. *Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004)*, 2004,
[61] Cai, L. and Hofmann, T., Text categorization by boosting automatically extracted concepts. *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, 2003, pp.182-189 (ACM Press, New York, NY, USA)
[62] Baker, L.D. and McCallum, A.K., Distributional clustering of words for text classification. *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, 1998, pp.96-103 (ACM Press, New York, NY, USA)
[63] Lewis, D.D., Representation and learning in information retrieval. *Computer Science*, 1992, (University of Massachusetts Amherst, USA)
[64] Loh, S., Wives, L.K. and Oliveira, J.P.M.d., Concept-based knowledge discovery in texts extracted from the Web, *ACM SIGKDD Explorations Newsletter*, 2000, 2(1), 29-39
[65] Liu, Y., Loh, H.T. and Tor, S.B., Building a Document Corpus for Manufacturing Knowledge Retrieval. *Proceedings of the Singapore MIT Alliance Symposium*, 2004,
[66] Dumais, S. and Chen, H., Hierarchical classification of Web content. *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR2000)*, 2000, pp.256 - 263
[67] Yang, Y., An evaluation of statistical approaches to text categorization, *Information Retrieval*, 1999, 1(1-2), 69-90