July 2008

DERIVING TECHNOLOGY ROADMAPS WITH TECH MINING TECHNIQUES

Fu-ren Lin  
*National Tsing Hua University, frlin@mx.nthu.edu.tw*

Chih-ping Wei  
*National Tsing Hua University, cpwei@mx.nthu.edu.tw*

Yu-kai Lin  
*National Tsing Hua University, ukailin@gmail.com*

Yu-liang Shyu  
*Institute for Information Industry, williams@iii.org.tw*

Follow this and additional works at: [http://aisel.aisnet.org/pacis2008](http://aisel.aisnet.org/pacis2008)

Recommended Citation  
Lin, Fu-ren; Wei, Chih-ping; Lin, Yu-kai; and Shyu, Yu-liang, "DERIVING TECHNOLOGY ROADMAPS WITH TECH MINING TECHNIQUES" (2008). *PACIS 2008 Proceedings*. 255.  
[http://aisel.aisnet.org/pacis2008/255](http://aisel.aisnet.org/pacis2008/255)
DERIVING TECHNOLOGY ROADMAPS WITH TECH MINING TECHNIQUES

Fu-ren Lin, Chih-ping Wei, Yu-kai Lin, National Tsing Hua University, Taiwan
fmlin@mx.nthu.edu.tw, cpwei@mx.nthu.edu.tw, ukailin@gmail.com;
Yu-liang Shyu, Institute for Information Industry, Taiwan, williams@iii.org.tw

Abstract

Technology monitoring has been a knowledge intensive and time-consuming task for IT managers or domain experts. Tech mining techniques can be used to mitigate these efforts. This paper proposes a technology monitoring framework based on tech mining techniques to facilitate the derivation of information and communication technology (ICT) roadmaps. With this framework, a tech mining engine is able to allocate the most relevant documents which describe a category of technologies. Domain experts were participated in a scan meeting to verify the generated roadmaps based on the selected cluster of documents. The draft roadmaps can be further articulated with domain experts’ judgment for technology forecasting and assessment.

Keywords: tech mining, technology monitoring, technology roadmap, scan meeting.
1 INTRODUCTION

Technology monitoring is important not only for the academic field but also for business and governments since the monitoring and forecasting of the evolution of technologies is essential for strategy setting and policy making. Monitoring is, according to Coates and Coates (1986), “to watch, observe, check, and keep up with developments, usually in a well-defined area of interest for a very specific purpose.” Porter and Datampel (1995) further pointed out that “monitoring is akin to ‘environmental scanning’ and ‘issues management’ – efforts to identify emerging developments likely to affect an organization over the coming few years.” However, with the exploding growth rate of online resources, the huge volume of documents made it hard to monitor the evolution of technologies. Hundreds of thousands of journal papers have been published or countless patents have been applied every year. The investigation of technology evolution highly relies on domain expertise. Experts have to involve the process from the beginning to the end which makes it slow and expensive. Moreover, it makes the automation of technology monitoring more imperative as the data volume grows enormously huge. Thus the major objective of this research is to apply tech mining techniques to build a tool that automatically performs technology monitoring, and facilitate domain experts to derive technology roadmaps.

Tech mining uses the methods of data/text mining to discover knowledge of specific domains of science and technology by exploring a huge amount of contents from full-text document corpora. Kongthon (2004) defines “text mining as the process that exploits large text collections to obtain valid, potentially useful and ultimately understandable knowledge which helps identify technology infrastructure and discovers research activities, etc.” Generally speaking, text mining can be divided into five major categories: document retrieval, data extraction, data processing, data analysis, and data visualization (Losiewicz, Oard, & Kostoff, 2000). The general issues targeted by tech mining include (1) what science and technology are being done globally, (2) who are major players in corresponding technical domains, (3) what are the levels of effort, (4) what are the relations among major thrust areas, (5) what are the relations between major thrust areas and supporting areas, (6) what are the promising directions for new research, and (7) what are the innovations and discoveries.

This paper is organized as follows. Section 2 reviews related literatures in technology monitoring and tech mining techniques. We propose the technology monitoring framework in Section 3. Section 4 demonstrates the architecture of tech mining engine with implemented techniques. Section 5 elaborates the findings from the implemented system, and Section 6 concludes this study.

2 LITERATURE REVIEW

Many related studies focus on extracting information from scattered resources to help understand science development and technology management. The most straightforward approach is to use clustering algorithms to group similar data. Teichert and Mittermayer (2002) used k-nearest neighbor and centroid-vector algorithms to classify patents into predefined categories. In their experiments, both algorithms got approximately 80% correct assignment comparing with experts’ assignments. The overall accuracy even boosts up to 90% combining both algorithms. The result shows that the automated text categorization can help significantly reduce the workload of experts, and then facilitate new techniques for new technologies.

Besides, Maynard et al. (2005) used an ontology-based framework to extract information for market monitoring and technology watch. They used GATE as a web service to implement in the h-TechSight portal. To evolve existing ontologies automatically with GATE in specific domains, users can monitor the domain-specific information. In the experiment, the application had been tested in the
employment sector with excellent results, and had been successfully ported to other genres of text such as news items and company reports. Wang et al. (2006) described GATE (Cunningham, 2002) as an architecture, a framework, and a development environment for LE (Language Engineering). As an architecture, it defines the organization of an LE system and the assignment of responsibilities to different components, and ensures that the component interactions satisfy the system requirements. GATE uses JAPE, a Java Annotation Patterns Engine, to recognize regular expressions in annotations on documents. JAPE provides a finite state transduction over annotations based on regular expressions. Regular expressions are applied to character strings, a simple linear sequence of items, but here these will be applied to much more complicated data structure. The result is that in certain cases the matching process is non-deterministic; that is, when there is structure in the graph being matched that requires more than the power of a regular automaton to recognize, JAPE chooses an alternative arbitrarily.

Kongthon (2004) proposed a framework based on text mining techniques to discover knowledge from a large amount of electronic text sources. It helps dig out helpful knowledge for research and development (R&D) programs. A novel text association rule mining algorithm has been proposed for gathering related concepts among text data. Two algorithms based on text association mining, called tree-structured networks and concept-grouping, have been implemented to apply to Thai science & technology (S&T) publication abstracts with the objective of improving R&D management. The results of their study can help support strategic decision-making on the direction of S&T programs in Thailand.

Emerging trend detection (ETD) plays an important role in technology monitoring. Emerging trend detection is a topic area that is growing in interest and utility over time. Kontostathis (2003) indicated that the knowledge on emerging trends is particularly important for individuals and companies to monitor a particular field or business. TimeMines (Swan & Jensen, 2000) is an automated system which generates timelines from date-tagged free text corpora. TimeMines detects, ranks and groups semantic features based on their statistical properties, and use these features to discover sets of related stories that deal with a single topic. Similarly, ThemeRiver Visualization (Havre, Hetzler, & Nowell, 2000) depicts thematic variations over time within a large collection of documents. The thematic changes are shown in the context of a time line and corresponding external events. The temporal thematic change within a context framework helps a user recognize patterns that suggest relationships or trends. Mei and Zhai (2005) proposed a temporal text mining task to discover and summarize the evolutionary patterns of themes in a text stream. A general probabilistic method has been proposed for solving the temporal problem through (1) discovering latent themes from text, (2) constructing an evolution graph of themes, and (3) analyzing life cycles of themes.

The process of retrieving information from documents comprises four phrases: data collection, data pre-processing, data analysis, and data visualization, where vector space model (Salton, 1989) is used to represent documents and measure similarity. To analyze documents, traditional data mining techniques, such as association-rule, classification, and clustering, can be useful tools. Ghani and Fano (2002) and Pierre (2002) use association-rule approach to uncover implicit rules in business domains. While patterns of observed targets are available and reliable, classification approaches leverage those features to build document classifiers. Ghanem et al. (2003) and Liu et al. (2004) applied text classification on biomedical areas. On the other hand, Hoitho, Maedche, and Staab (2003) and Beil, Ester, and Xu (2002) adopted clustering algorithms for text clustering. To improve clustering quality, Hoitho et al. further incorporated ontology into vectors by adding or replacing terms by concepts. The ontology-based approach is becoming important due to the fact that term relationships cannot be found lexically from time to time.

Although there are many related work, most of them could not entirely fulfill users' anticipations in answering the general issues targeted by tech mining identified in Introduction. Table 1 summarizes the functionalities of aforementioned related works and compares them with this study in the seven issues.
### Table 1. System capability of related work and this study

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) what science &amp; technology is being done globally</td>
<td>V</td>
<td>V</td>
<td>V</td>
<td>V</td>
</tr>
<tr>
<td>(2) who are major players in corresponding technical domains</td>
<td>X</td>
<td>V</td>
<td>X</td>
<td>V</td>
</tr>
<tr>
<td>(3) what are the levels of effort,</td>
<td>P</td>
<td>V</td>
<td>X</td>
<td>V</td>
</tr>
<tr>
<td>(4) what are the relationships among major thrust areas</td>
<td>P</td>
<td>P</td>
<td>X</td>
<td>P</td>
</tr>
<tr>
<td>(5) what are the relationships between major thrust areas and supporting areas</td>
<td>X</td>
<td>P</td>
<td>X</td>
<td>P</td>
</tr>
<tr>
<td>(6) what are the promising directions for new research</td>
<td>P</td>
<td>X</td>
<td>X</td>
<td>P</td>
</tr>
<tr>
<td>(7) what are the innovations and discoveries</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>V</td>
</tr>
</tbody>
</table>

*(V: support; P: partially support; X: non-support)*

### 3 PROPOSED FRAMEWORK

Considering the needs of automatic technology monitoring and the prospect of tech mining techniques, we propose a technology monitoring framework shown in Figure 1 that provides an end-to-end solution. First, target technologies are selected and sources of information are determined by domain experts. Tech mining engine will be developed and used for extracting relations among features pertaining to each target technology and for generating visualized results. Scan meetings composed of domain experts will be held to discuss these findings and then conclude technology monitoring and forecasting reports.

**Figure 2. The proposed technology monitoring framework**

3.1 **Technology selection**

In the pre-processing phase, the first step is to decide which technologies to monitor. Domain experts have to allocate sources of information which can be used for discovering emerging technologies. For different domains of technology, the sources of data collection will be different as well. For instance, some technologies are highly practice-oriented ones, such as Web 2.0, Really Simple Syndication (RSS), Business Intelligence (BI), to name a few, which are more likely to be monitored through news, magazines, and industry reports. On the other hand, some are more academic such as communication system performance, energy management, or optimization methods, which should be observed in journal papers. Moreover, a variety of technologies lie in between of the spectrum of interest.

3.2 **Tech mining engine**

Figure 2 illustrates four modules, *i.e.*, data collection, data pre-processing, data analysis, and data visualization in a tech mining engine, which automates most processes that are used to be expensive and time-consuming.
Due to the fact that different data sources share no common format, the data collection module is virtually a set of agents; each agent is in charge of a data source. They will crawl the data sources automatically, and collect data related to the selected technologies in the previous stage.

Most of the collected data are documents. Therefore, information should be extracted from them. The data pre-processing module deals with extracting key terms which may represent technologies, institutes, researchers, etc. from those collected documents. Besides information extraction, the data has to be cleaned thoroughly and translated into proper formats in order to exploit them.

The data analysis module discovers the implicit relations among entities along the time line, such as technologies, researchers, and institutions by processing data generated from the pre-processing stage. Both supervised and unsupervised learning algorithms can be applied to data analysis. Only, unsupervised learning algorithms such as clustering methods (Baeza-Yates and Ribeiro-Neto, 1999) seem to be more adaptive and reasonable to investigate emerging trends because of the rapid-changing atmosphere of technology development and the difficulty to define patterns beforehand.

The visualization module is to display relations discovered in the data analysis module. An interactive interface is useful for users’ impromptu browsing and queries. The results can be presented quantitatively and qualitatively. The visualization should also contain reference materials to facilitate users to investigate the details.

### 3.3 Scan meetings

The outputs of the previous phase can be further articulated with domain experts’ judgment for technology forecasting and assessment. A scan meeting aims at explaining and discussing the outputs from the proposed tech mining engine in order to conclude final technology monitoring and forecasting reports. Due to the diversity of the monitored technologies, the participant experts of the scan meeting will come from different domains. The outputs of the tech mining engine present preliminary relations among technologies, and between technologies and corporations, and we still need experts from different areas to enrich the profundity of the final reports. We can reduce a great deal of efforts and produce reliable information by combining the outcome of the tech mining engine and the expertise from the scan meeting.

### 4 IMPLEMENTATION

In this study, we developed a platform and used information and communication technology (ICT) as a scenario to exemplify the proposed technology monitoring framework. Figure 3 demonstrates main elements in the platform, including data sources, a sequence of processes, and implemented functions in each module of the tech mining engine.
First, we prepared a set of seed terms which are interesting to users. These terms mainly are Internet and communication technologies or applications, such as CDMA, compression, ERP, grid, IPTV, social computing, streaming, to name just a few (see Appendix A for the full list of seed terms).

We then developed a tech mining engine that automates most of the technology monitoring tasks. It will collect data from three major sources: ISI Web of Knowledge, Gartner.com, and Forrester.com. ISI Web of Knowledge maintains a large collection of SCI journal papers which serve our needs for investigating academic research. Among hundreds of subject areas, we consider subject area names that prefix “compu”, “commu”, or “info” as ICT related. Accordingly, the data collection module sends queries to ISI web site, limits results in subject areas of ICT within the past ten years, and save them. Meanwhile, Gartner and Forrester can both deliver technology-related industrial and practical information. In the end, 46,300 journal papers and 5,931 industry reports were collected. To filter and rank documents in the following phrases, we further incorporate paper citation numbers from Google Scholar and journal impact factor which is also available on ISI Web of Knowledge.

The data pre-processing module of the tech mining engine identifies and extracts document profiles, such as authors and publication time, and terms that are related to technologies. The module also purifies data by handling stop words, stemming, mapping acronyms, and filtering insignificant terms. The filtering considers two criteria: threshold and significance. We heuristically set term frequency an upper bound 2000 and a lower bound 20 to exclude terms that are too common or too rare. The significance, on the other hand, is determined by average slopes of appearance frequencies of entities in the past 10 years; the higher slope value signifies higher potential. We filter ten percent least significant terms in the experiment. 210,983 recognized technology terms were identified by the data pre-processing module, and 540 retained afterward.

We then used TF-IDF and cosine similarity to build vectors for these retained 540 terms from the previous module, and applied Hierarchical Agglomerative Clustering (HAC) (Kaufman and Rousseuw, 1990) to recognizing implicit relations among technologies. To determine the best number of clusters, which is to maximize intra-cluster similarity and to minimize inter-cluster similarity, we utilized the min-max cut method proposed by Chuang and Chien (2005). Based on their proposed equations, the best number of clusters is that with the lowest min-max value. Figure 4 shows that 28 is the best number of clusters in the experiment.
The visualization module adopts perfuse (Heer, Card, & Landay, 2005), an open source, java based toolkit. It is characterized for its rich visualization library and built-in highly interactive interfaces. Currently, the developed functions on the graphic user interface of the tech mining engine are exemplified in Figure 5. These functions include (1) trend viewer that demonstrates the popularity of technologies along the time line, (2) expert networks which indicate who specializes on a particular technical domain or what domains interest a particular expert, (3) networks of related technologies that identifies the relations among different technologies, (4) ranking clusters and articles according to their significance on academic or practice, and (5) article viewer that presents important materials to read.

Among clusters generated from the tech mining engine, we chose a cluster, which is related to wireless communication (see Appendix B for the complete list of the cluster), to derive a technology roadmap. We then hosted a scan meeting joined by industry analysts and academic researchers specialized in the selected domains to validate the draft roadmap. Based on their feedbacks, the final modified technology roadmap is shown in Figure 6.
We categorized the reference materials into three sections—media applications, infrastructure, and protocol standards. The media applications are mainly from reference materials of three terms—cellular network, downlink, and channel capacity—from industrial reports, where mobile TV broadcasting, IPTV, and VOIP WWAN are popular topics among all mentioned technical applications. The infrastructure can be referred from both industrial reports and journal papers. Wireless mesh network seems to be a promising network infrastructure. 802.11x, WiMA, and WiMax can be its underlying transmission model. These reports also indicate that UWB (ultra-wideband) is an important technology and will become widely used in PAN (personal area network) in five years. The protocol part refers to academic researches, where HSDPA (high speed downlink packet access) is highly emphasized. Although many experts predict that 4G will be the mainstream in five years, CDMA 2000 and WCDMA will still be the standards in the transition.

Another finding provides the evidence to support a widely recognized issue: chasms between academic research and practice to the public. We investigated this phenomenon by calculating popularities of every cluster in both academic and industrial domains. The popularity is determined by the frequency of term appearance. However, due to the fact that they are available in one year, we only consider the industry environment of 2007. Figure 7 demonstrates the comparison of these two domains, and the ramified ranking proves the phenomenon. Nevertheless, it is still hard to tell which one leads the other without comparing long term industrial reports and academic research.
5 CONCLUSION AND FUTURE DIRECTIONS

Technology monitoring is a continuous observation on emerging technologies. It, traditionally, requires domain experts’ time and effort on collecting, organizing, and analyzing data. The rapid changing environment of information and communication technology industry makes technology monitoring even harder. This study proposes a technology monitoring framework and uses tech mining techniques to develop a technology monitoring platform to facilitate and automate the process of data collection, data purification, data analysis, and data visualization. We take its output to derive a technology roadmap and verify it by a group of domain experts in a scan meeting.

We have learned two things from this framework and its implementations. First, consuming the huge amount of documents, automated tools do significantly help technology monitoring. We notice that a small number of seed terms incur tens of thousands related articles. Therefore, it is not difficult to imagine the tremendous cost of monitoring them manually. Second, we also witness the shrink, followed by the expansion in the beginning, on the amount of processed data in the framework. Scan meetings can complement the tech mining engine on articulating its outputs.

However, our draft roadmap is somehow out of accord with the experts’ mind set. The data driven approach on data analysis and data visualization seems to be difficult to represent semantic relations of technologies and their hierarchical dependency, which encourages an ontology-based approach as the direction of future works.

References


APPENDIX A

Seed terms:
APPENDIX B

Terms of the cluster used to derive the roadmap:

SDMA (space division multiple access), adaptive array, smart antenna, BF (bloom filter), adaptive algorithm, adaptive antenna array, AA (antenna array), DSCDMA (direct sequence code division multiple access), code acquisition, acquisition, outage probability, SIR (step impedance resonator), TPC (turbo product code), antenna diversity, multiaccess communication, TH (time hop), PPM (pulse position modulation), iterative decoding, convolutional code, turbo code, SE (spectral efficiency), channel capacity, blind channel estimation, diversity method, rake receiver, multipath channel, multipath, MUI (multiuser interference), multicarrier cdma, TR (transmitted reference), NBI (narrow band interference), DFE (decision feedback equalizer), SC (single carrier), FDE (frequency domain equalization), equalization, FD (finite difference), CP (cyclic prefix), ISI (inter-symbol interference), ICI (inter-carrier interference), synchronization, frequency offset, CFO (carrier frequency offset), TSTD (transmit diversity), STBC (space time block code), DF (decision feedback), MAI (multiple access interference), MMSE (minimum mean square error), MUD (multiuser detector), multipath fading channel, frequency selective fading, fading channel, rayleigh fading, correlation, wireless communication, communication system, ST (space time), MIMO, FDM (frequency division multiplexing), CCI (cochannel interference), land mobile radio cellular system, multiuser channel, iterative method, IC (integrated circuit), PIC (particle in cell), SiC (silicon carbide), MRC (maximal ratio combining), SD (service discovery), signal detection, computational complexity, ML (maximum likelihood), MLD (maximum likelihood detection), AMC (adaptive modulation and coding), AM (adaptive modulation), packet scheduling, downlink, HSDPA (high speed downlink packet access), SA (system architecture), OFDMA (orthogonal frequency division multiple access), power allocation, RRM (radio resource management), resource allocation, resource management, broadband, TDMA (time division multiple access), mac protocol, LP (linear programming), optimization method, DP (dynamic programming), cellular system, soft handoff, handoff, cellular network, CAC (call admission control), admission control