Analysis of Technology Trends Based on Big Data

Aviv Segev Chihoon Jung Sukhwan Jung
Department of Knowledge Service Engineering
KAIST
Daejeon, Korea
aviv@kaist.edu chihoon.jung@kaist.ac.kr raphael@kaist.ac.kr

Abstract — **The paper suggests a method for analyzing technology trends. The process, which investigates development of technologies over time, identifies main technologies displaying the fastest growth compared to greater influence of new inventions. The method analyzes term frequency and change over time of technological terms in patents to identify the prior technologies that lead to a new technology and detects technologies that have the biggest impact. The analysis was performed on 4,354,054 patents from the US Patent Office dating from 1975 until today. Some correlation is displayed between technology trends and future US stock market performance.**

I. INTRODUCTION

The topic of identifying new technologies has implementations in the area of stock prediction, technology venture funds, and government research investment planning. The current work presents a method for analyzing technology trends and identifying the cause and effect of a given technology. The method is based on temporal term frequency analysis compared with first and second derivatives of change of similar technologies. This perspective presents both growth and influence of a specific technology in a specific time period. These technologies are compared to identify cause and effect of specific technologies and technology trends that have major influence on innovation over time.

II. RELATED WORK

Previous work in Information Retrieval (IR) has targeted patent documents. During the NTCIR (NII Test Collection for IR Systems) Workshops, in patent retrieval tasks a test collection of patent documents was produced and used to evaluate a number of participating IR systems. One task analyzed geographic and temporal information retrieval [1], with the focus to perform searches with geographic and temporal constraints. The data collections (Japanese and English news stories) combined geographical IR with time based search to find specific events in a multilingual collection. Other attempts at patent classification focused on cross-lingual link discovery (CLLD) [2], which sought to automatically find potential links between documents in different languages. The goal of this NCTIR task was to create a reusable resource for evaluating automated CLLD approaches. The goal of the task was to build and refine systems for automated link discovery. The task focused on linking between English source documents and Chinese, Korean, and Japanese target documents. Currently, the NTCIR tasks aim at machine translation of sentences and claims from Chinese to English, Japanese to English, and English to Japanese [3].

The Workshop of Cross-Language Evaluation Forum (CLEF 2009) [4] gave separate topic sets for the language tasks, when the document language of the topics was English, German, and French. CLEF-IP included Prior Art Candidate Search task (PAC) and Classification task (CLS). Participants in the PAC task were asked to return documents in the corpus that could constitute prior art for a given topic patent. Participants in the CLS task were given patent documents that had to be classified using the International Patent Classification codes. In addition, evaluations were performed on chemical datasets in chemical IR in general and in chemical patent IR in particular. A chemical IR track in TREC (TREC-CHEM) [5] addressed the challenges in chemical and patent IR.

Previous work analyzes automatic patent retrieval, while this work describes a method that involves a manual decision process assisted by an automatic suggestion of relevant concepts related to patent technology evolution over time.

III. TECHNOLOGY TRENDS ANALYSIS METHOD

The technology trends analysis method is based on analyzing a large data set of technology-based documents such as patents. The data set is assumed to be organized sequentially by date of issue. The method includes identification of the main terms related to a given technology, the next step involves extraction of the sequential graph describing the frequency of the terms, followed by an elimination of graphs with different behavior, and finally identification of graphs with closest delta distance that represent the cause and effect of the analyzed technology. In addition, the first and second derivatives are evaluated to analyze the magnitude of the effect.

The analysis was performed on 4,354,054 patents from the US Patent Office dating from 1975 until today. An example of *email* technology is displayed in Figure 1. The method allows an identification of contributing technologies that led to the fast growth of the *email* technology.

A. Extracting Related Terms

The first step identifies all the terms related to the technology analyzed. A method to extract the relevant terms can include extracting all the linked terms that appear in the technology term description in Wikipedia. The extracted term list can be filtered, and additional terms can be added manually.

B. Extracting Relevant Technologies

The second step involves extracting values that represent term frequency in a large data set of documents that can represent the different technologies. An example of such data sets can be patents or research publications. The term frequency uses simple keyword search in either the subject, abstract, full description of the document, or all of these options. The time slot being analyzed usually is a year, since

978-0-7695-5006-0/13 $26.00 © 2013 IEEE
DOI 10.1109/BigData.Congress.2013.65

IEEE computer society

smaller time slots can entail high incidents of seasonal noise. The term frequency has to be weighted, since the extraction searches for an increase in term frequency rather than just elevated values. The weight method analyzed used $max_j\ (tf_j)$ value on all technology term frequencies. Other terms such as $(tf_i - tf_{i-1})/max_j\ (tf_j)$ were also evaluated.

The elimination process includes identifying all the graphs that do not represent exponential growth of a new technology. The following types of regression functions were analyzed to identify the best fitting function for technology growth including linear, quadric, cubic, quadratic, exponential, and mixed functions. The best matching function based on a predefined set of samples of existing technologies was an exponential regression and the values selected as coefficients were based on the average values of the sample technology functions: $y= 0.055558046* 1.160450815\text{^}x -0.084088217$

For all technologies, the coefficient of determination R^2 was calculated as the square of the sample correlation coefficient between the outcomes and their predicted values in the matching function y. If the value of $R^2<0.94$, then the technology was discarded as not representing new technology exponential growth. An example of the results of term frequency of related terms to *email* technology is presented in Figure 1 (top).

C. Cause-Effect and Impact of Technologies

Once all the technologies with exponential growth have been identified, the next step includes classifying technologies by cause, effect, and how much impact each technology has. The coefficient of determination is used again to identify the distance between the technology being analyzed and all other technologies. A similar process is used based on predefined Δt time difference. The Δt represents time for one technology to be influenced by the other. If the majority of the data samples are before a specific technology, then the current technology is a predictive cause of the specific future technology. If the majority of samples are after the technology, then the current technology is a cause of the new technology, or an effect of the analyzed technology.

The first derivative, $y_i'=\Delta y_i/\Delta t$ describes the extent of the growth of a specific technology. Comparing technologies with similar graph behavior (growth at specific time periods) shows the main technologies that contributed to a specific trend. For example, the extended use of email can be attributed mainly to technologies such as computer network, internet, server, and mobile, and to a less extent digital message, as displayed in Figure 1 (middle).

The second derivative $y_i''=\Delta(\Delta y_i/\Delta t)/\Delta t$ displays the acceleration of change of technology trends in the market (Figure 1 – bottom). The results display 4 major time periods with a drop in stock market performance that might be attributed to technology: 1999 (2000 dot-com bubble), 2007 (2008 financial crisis), 2011 (August 2011 stock market fall). Although the stock market prediction shows promising results, a few issues should be considered. The year 2005 also shows a drop in technology patents. However, the stock market did not show a similar fall the year after. A more important issue is that the data used were granted patents. In other words, the information is based on patents that were submitted 3-4 years before to the patent office. Currently full scale experiments are underway to verify the initial results. Further research includes analyzing smaller time slots and isolating noise.

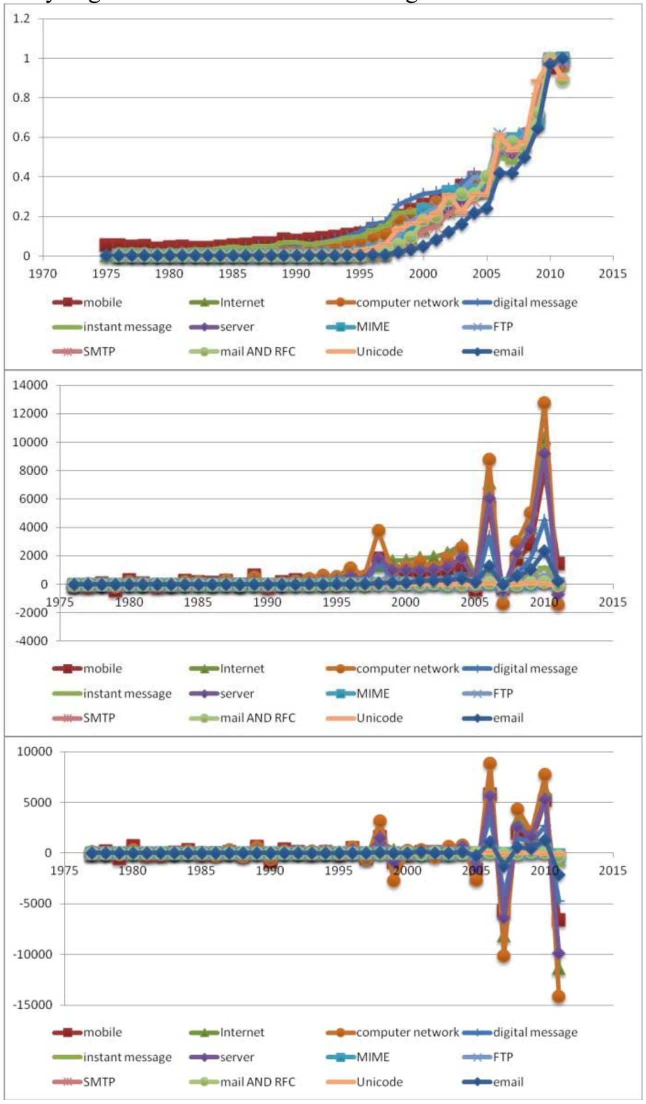

Figure 1. E-Mail Trend Technologies (Top) Impacting Technologies (Middle) Acceleration/Decceleration of Technology Trends (Bottom)

References

[1] F. Gey, R. Larson, J. Machado, M. Yoshio, NTCIR9-GeoTime Overview - Evaluating Geographic and Temporal Search: Round 2, Proceedings of NTCIR-9 Workshop, 2011, pp. 9-17.

[2] L. Tang, S. Geva, A. Trotman, Y. Xu, K. Itakura, Overview of the NTCIR-9 Crosslink Task: Cross-lingual Link Discovery, Proceedings of NTCIR-9 Workshop, 2011, pp. 437-463.

[3] I. Goto, B. Lu, K. Chow, E. Sumita, B. Tsou, Overview of the Patent Machine Translation Task at the NTCIR-9 Workshop, Proceedings of NTCIR-9 Workshop, 2011, pp. 559-578.

[4] G. Roda, J. Tait, F. Piroi, V. Zenz, CLEF-IP 2009: Retrieval Experiments in the Intellectual Property Domain, in: Proceedings of the Workshop of the Cross-Language Evaluation Forum (CLEF 2009), 2010, pp. 385-409.

[5] M. Lupu, J. Huang, J. Zhu, J. Tait, TREC-CHEM: Large Scale Chemical Information Retrieval Evaluation at Trec, SIGIR Forum 43 (2), 2009, pp. 63-70.