Temporal Term Frequency Analysis of Technology

Aviv Segev
Department of Knowledge Service Engineering
KAIST
Daejeon, Korea
aviv@kaist.edu

***Abstract* — The paper suggests a method for analyzing cause and effect of technology. The process can be identified as a flowchart of technologies over time. The method analyzes term frequency of technological terms in patents to identify the prior technologies that lead to a new technology and the identified technology outcome. The analysis was performed on 4,354,054 patents from the US Patent Office from 1975 until today.**

I. INTRODUCTION

The problem 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 and identification of similar technologies that present exponential growth. These technologies are compared to the analyzed technology to identify cause and effect according to the prediction ability of each technology based on their coefficient of determination value over a delta time difference from the original technology.

II. RELEATED WORK

Previous work in Information Retrieval (IR) has targeted patent documents. During the NTCIR Workshops [1], [2] a patent retrieval task was organized in which a test collection of patent documents was produced and used to evaluate a number of participating IR systems. In the NTCIR-3 Patent Retrieval Task, participant groups were required to submit a list of relevant patent documents in response to a search topic consisting of a newspaper article and a supplementary description. Search topics were in four languages. All topics were initially written in Japanese and were manually translated into English, Korean, and traditional or simplified Chinese. In NTCIR-4 the search topic files were Japanese patent applications that were rejected by the Japanese Patent Office. The English patent abstracts were human translations of the Japanese patent abstracts. Currently, the NTCIR tasks aim at machine translation of sentences and claims from Japanese to English. Other work analyzed Japanese-English cross-language patent retrieval using Kernel Canonical Correlation Analysis (KCCA), a method of correlating linear relationships between two variables in the kernel defined by feature spaces [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 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 TEMPORAL ANALAYSIS METHOD

The technology temporal 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 identifying the main terms related to a given, the next step involves extracting 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.

A. Extracting Related Terms

The first step identifies all the terms related to the technology being analyzed. A method to extract the relevant terms can include extracting all of 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 All Graphs

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 involves a year since smaller time slots can entail high incidents 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. An example of the results of term frequency of related terms to *email* technology is presented in Figure 1 (top).

C. Elimination Process

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. The best matching function based on a predefined set of sample of existing technologies was an exponential regression and the values selected as coefficients were based on the average values of the sample technology functions:

$$y= 0.055558046* 1.160450815^x -0.084088217$$

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.

D. Graph Distance

Once all of the exponential growth technologies have been identified, the next step includes classifying technologies that are cause, effect, or non-related to the technology being analyzed. 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 the possible prediction time of one technology affected by the other. The last step identifies the number of data samples that appear before and after the analyzed technology. If the majority of the data samples are before, then the current technology is a predictive cause of the current technology (Figure 1 - middle). 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 (Figure 1 - bottom).

IV. Technology Temporal Analysis Experiences

The analysis was performed on 4,354,054 patents from the US Patent Office from 1975 until today. An example of *email* technology is displayed in Figure 1. The method allows an identification of contributing technologies which led to the fast growth of the *email* technology. In addition, the method enables the elimination of possible irrelevant technologies that existed at the time but did not directly contribute directly to the analyzed technology. Additional work is currently being performed to create an ongoing flow chart of all technologies that presented exponential growth and their contribution to other new technologies.

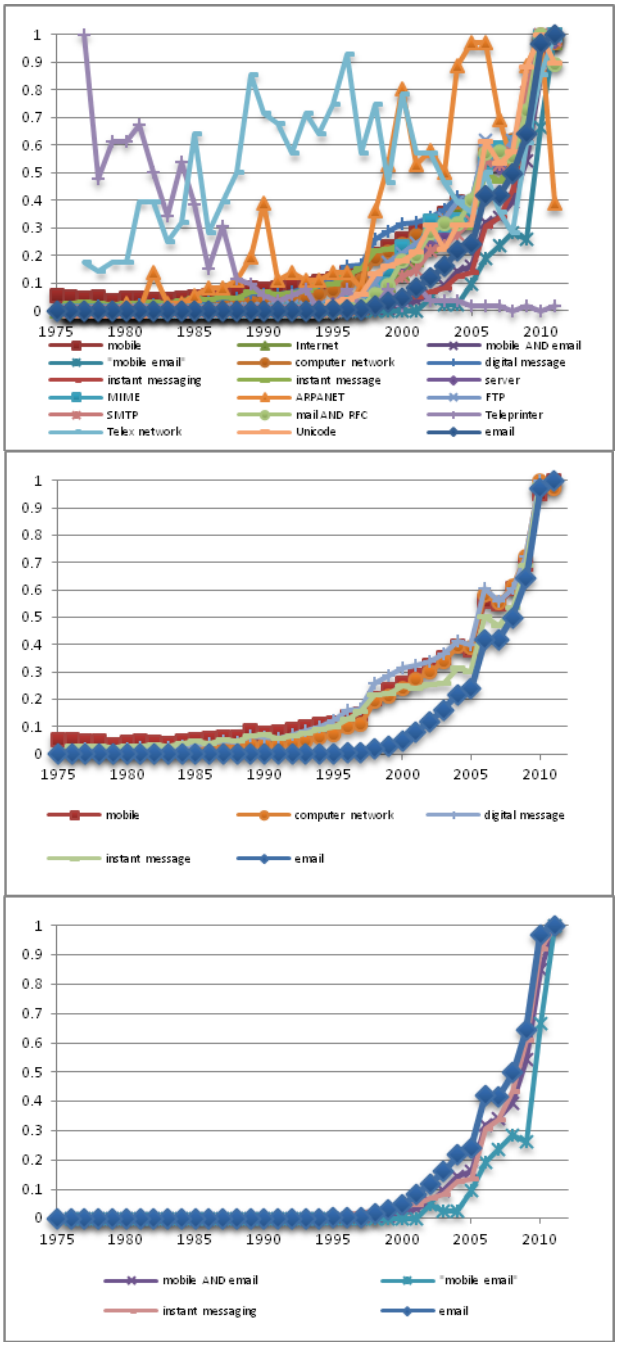

Figure 1. E-Mail Temporal Technologies (Top) Technology Identified Cause (Middle) and Technology Identified Effects (Bottom)

References

[1] M. Iwayama, A. Fujii, N. Kando, Y. Marukawa, Evaluating patent retrieval in the third NTCIR workshop, Information Processing and Management 42 (2006) 207-221.

[2] A. Fujii, M. Iwayama, N. Kando, The patent retrieval task in the fourth NTCIR workshop, in: Proceedings of the SIGIR-04, 2004, pp. 560-561.

[3] Y. Li, J. Shawe-Taylor, Advanced learning algorithms for cross-language patent retrieval and classification, Information Processing and Management 43 (5) (2007) 1183-1199.

[4] G. Roda, J. Tait, F. Piroi, V. Zenz, CLEF-IP 2009: retrieval experiments in the intellectual property domain, in: Proceedings of the 10th 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) 63-70.