Computers & Industrial Engineering 115 (2018) 532–542

Contents lists available at ScienceDirect

Computers & Industrial Engineering

journal homepage: www.elsevier.com/locate/caie

Semi-automatic extraction of technological causality from patents

Hongbin Kim, Junegak Joung, Kwangsoo Kim*

Department of Industrial and Management Engineering, Pohang University of Science and Technology, San 31, Hyoja-dong, Nam-gu, Pohang, Kyungbuk 790-784, Republic of Korea

ARTICLE INFO

Keywords:
Patent analysis
Cause-and-effect relationship
Causal relation
Causal patterns
Technology analogy

ABSTRACT

The goal of this study is to suggest a method to extract technological causalities from patents, which are formal documents that include a large amount and large variety of information about technology. The core of patents is composed of both inventive principles to solve problems, and purposes that the invention achieves by solving them. The principles and purposes can be understood as a concept of technological causality which is reusable knowledge as technological analogy. Because reading and understanding patent documents that generally consist of dozens of pages and have difficult and profound statement of technologies is hard even for technology experts, a method to extract technological causalities is needed. As a solution, this paper proposed a method to extract technological data from patents, to identify technological causes and effect relation from the extracted data and to calculate the representativeness of technological causes and effect. Based on this study, technology experts can be given a list of ranked alternatives for technological causes and effect. This study helps to analyze patent, and it finally contributes to new product development and technology opportunity discovery. To achieve the objectives, the proposed method included the characteristics of patents that are structured documents consisting of various particular fields that have each different contents and importance. And natural language processing technology is adopted to automatically extract meaningful data and to perform linguistic processing. The implementation and case study of the proposed method demonstrated how a prototype system can be developed and utilized.

1. Introduction

Patent analysis has been in the limelight to obtain meaningful technological insights from patents which are formal documents containing voluminous and various technology information, and are even free to use and easily accessible. Patent analysis can be divided into two categories depending on what kind of data is used. Firstly, some studies use bibliographic information such as assignee, references (forward reference and backward reference), regal status, designated state or International Patent Code with the various purposes of estimating value of patent, developing indicators, and analyzing knowledge flows (Harhoff, Scherer, & Vopel, 2003; Jaffe & Trajtenberg, 1999; Ko, Yoon, & Seo, 2014; Narin, 1994; Trajtenberg, 1990). Secondly, other studies use contents of patents such as title, abstract, description and claims to exploit technological knowledge presented in patents. In these study, various data models such as keyword, Subject-Action-Object (SAO), and Property-Function (PF) are used to extract technological data. Because there are various kinds of data even in a single patent document, studies on patent analysis have focused on what kind of data can be utilized to discover meaningful technological insights.

However, we focused on what kind of contents is important in patent document not the kind of data. Even patent documents have various and voluminous data, the core of patent is about technology or invention. More minutely, an inventive principle to solve problems and a useful purpose of invention are the most important data in a patent. Once an invention is patented, it is guaranteed that it has novel principle and valuable purpose. And these can be understood as a concept of causality. An inventive principle is a technological cause of a patent, and a useful purpose is its technological effect (Kim & Kim, 2012). Since cause-and-effect relationship is one of the most basic ways of thinking, it facilitate understanding of phenomena. Therefore, a technological causality extracted from a patent can represent the core contents of it.

As a legal document, patent is given a right to exclude others from making, using or selling the invention (Organization, 2004). However, the legal right is assigned only to the claim field in patent documents which states constitutions of inventions. On the contrary, the elements of technological causalities, inventive principles and useful purposes, are not subject to the right of patent, because cause-and-effect relationship of technology is a kind of nature law or scientific theory which is not the scope of patentable subject matter. For example,

* Corresponding author at: Industrial & Management Engineering, POSTECH, Republic of Korea.
E-mail address: kskim@postech.ac.kr (K. Kim).

https://doi.org/10.1016/j.cie.2017.12.004
Received 2 July 2015; Received in revised form 7 September 2017; Accepted 4 December 2017
Available online 06 December 2017
0360-8352/

US7306002 utilized centrifugal force to clean wafer in semiconductor fabrication processing. The technological cause of the patent is to utilize centrifugal force and technological effect is to clean object, which are also found in the technology of vacuum cleaner or domestic spin dryer. As a result, extracting and defining a causality from a patent means that contents of patent technology are condensed as a most abstractive expression. And more importantly, extracted technological causalities can be reusable and applied to other domains.

Another advantage of technological causalities is that they can be connected building a network consisting of nodes (causes and effects) connected by links (causalities). When different technologies have either a same cause or an effect, they can be collected together with the same cause or effect as a center. For example, centrifugal force can be utilized not only to clean object but also to separate mixture or other purposes. And cleaning object can be achieved not only by centrifuge force but also by vacuum effect or other principles. Consequently, the technological causality network is where heterogeneous technologies can be connected with the concept of causality as a medium. The network can be utilized in two ways. Technology experts in new product development (NPD) can find various principles or solutions to solve problems by inferring causes, and entrepreneurs or decision makers can discover new applications or opportunities to utilize existing technologies. The result from the network is made up of information from not a single domain but diverse domains. The technological causality network can be served as a new way of patent search system compared to the previous patent search engine where users query with keywords and only can have results related to those keyword.

However, much research to identify causal information is not suitable for NPD or technology opportunity discovery (TOD) using patent analysis, and is not sufficiently detailed to allow identification of causal relationships. Altenberg (1984) and Nedjalkov and Silnickij (1973) used linguistic clues to identify causal relations, but the method is not appropriate for analysis of patent documents, which are composed of technical terms. Kontos and Sidiropoulou (1991) and Kaplan and Berry-Rogghe (1991) used hand-crafted rules and domain knowledge to extract causal relations, but crafting every rule based on domain knowledge is difficult manual work. Khoo, Kornfilt, Oddy, and Myaeng (1998) extracted many cause-and-effect relations from the Wall Street Journal, but their method cannot disclose representative cause-effect relations in a single article. To complement the manual work, Romacker and Schulz (2001) used a dependency parser to identify medical knowledge automatically, but their method is limited to the medical domain and focuses on avoiding duplicated knowledge; the method does not extract causal information. Fantoni, Apreda, Dell'Orletta, and Monge (2013) extracted function-behavior-state information similar to cause-and-effect relations from patents, but the method identifies all information related to various functional verbs and behaviors; as a result, the extracted information is very noisy, so users may have to devote substantial time to refining the results.

As a remedy for the current limitations, this paper suggests a semi-automatic method and system to identify technological causalities from patents. To achieve this goal, this study uses the Stanford dependency parser and rules to extract cause-and-effect relationships automatically. The proposed method can identify complex cause-and-effect relationships that include adjectives, adverbs, and verbs. Moreover, cause-and-effect relationships that represent a single patent can be identified by weighting causes and effects found in parts of the patent (title, abstract, and description). Therefore, this method helps to build a technological causality network efficiently, and contributes to NPD and TOD that use patent analysis.

The rest of this paper is organized as follows. Section 2 describes related work on the extraction of causalities from text. Section 3 proposes a semi-automatic method to extract technological causalities. Section 4 illustrates an implementation of the system and case study. Finally, Section 5 presents conclusions and future research.

2. Related work

The early studies to extract causal relations from text were focused on the definition of the linguistic patterns or cues to identify the causal relations. Altenberg (1984) defined four types of causal link and an extensive list of such linking words reconciled from several sources, including Greenbaum (1969), Halliday and Hasan (1976), and Quirk, Greenbaum, Leech, and Svartvik (1972): (1) Adverbial linkage (e.g., so, hence, therefore). (2) Prepositional linkage (e.g., because of, on account of). (3) Subordination linkage (e.g., because, as, since). (4) Clause-integrated linkage (e.g., that's why, the result was). Nedjalkov and Silnickij (1973) made multilingual causation studies and classified the causation verbs as the following categories: (1) Simple causatives - the linking verb refers only to the causal link, most of the time being synonymous with cause (e.g., Earthquakes generate tidal waves). (2) Resultative causatives - the linking verb refers to the causal link plus a part of the resulting situation (e.g., kill, melt, dry, break, drop). (3) Instrumental causatives – they express a part of the causing event as well as the result (e.g., poison, hang, punch, clean). The linguistic clues to extract causal information in the early stage were too small and limited to catch sophisticated causes and effects in technology fields. For example, causative adverbs such as therefore, hence, and conditionals like 'if… then…' construction are seldom used in patent documents. Also they were defined for normal natural language like news, books or webpages but not specialized for technological text.

And the early studies mostly depended on knowledge-based inferencing to extract and define causal relations. Selfridge, Daniell, and Simmons (1985) and Joskowicz, Ksiezyck, and Grishman (1989) developed software for an expert system that extracted causal information from short explanatory text. The problem was that when there was ambiguity about whether a causal relationship between two events is expressed in the text, the system used the domain knowledge to check whether a causal relationship between the events is possible. Kontos and Sidiropoulou (1991) and Kaplan and Berry-Rogghe (1991) used linguistic patterns to identify causal relations in scientific texts. However, substantial domain knowledge and knowledge-based inferencing were needed for identifying causal relations in the sample texts accurately. And the information for linguistic processing such as the grammar, the lexicon, and the patterns for identifying causal relations were hand-coded and were developed just to handle the sample texts used in the studies.

Khoo et al. (1998) investigated how cause-effect information can be extracted from text without knowledge-based inferencing and without full parsing of sentences. He made the set of linguistic patterns for identifying causal relationships, by adapting and modifying the rules or patterns of the several previous studies. Data used in this study was wall street journal covering a wide range of subject, not to deal with only narrow domains. However, the study concerned about the extraction of causal relations not about importance or representativeness of them. Therefore even if we have many causal relations from a single article, we don't know which causal relation is important or representative.

Girju and Mldovan (2002) presented a semi-automatic method of discovering generally applicable lexico-syntactic patterns that refer to the causal relation. In this study, the patterns were found automatically, but their validation was done semi-automatically in accordance with ambiguity of verb. Girju (2003) has evolved the research and later presented an inductive learning approach to the automatic discovery of lexical and semantic constraints necessary in the disambiguation of causal relations that are then used in question answering for QA system. The both studies focused on the only syntactic-patterns of the form < NP1 verb NP2 > , where the verb is causative verb since the form is one of the most frequent explicit intra-sentential pattern that can express causations. However, the causal relation is not limited to that form. The causal relation can be expressed in text in various ways. The problem of this situation is that Noun phrases (NPs) only can be defined as causes or effects. Causes and effect can be expressed with not only noun