

## Methodologies for Deriving Semantic Technical Information Using an Unstructured Patent Text Data

**Authors :** Jaehyung An, Sungjoo Lee

**Abstract :** Patent documents constitute an up-to-date and reliable source of knowledge for reflecting technological advance, so patent analysis has been widely used for identification of technological trends and formulation of technology strategies. But, identifying technological information from patent data entails some limitations such as, high cost, complexity, and inconsistency because it rely on the expert' knowledge. To overcome these limitations, researchers have applied to a quantitative analysis based on the keyword technique. By using this method, you can include a technological implication, particularly patent documents, or extract a keyword that indicates the important contents. However, it only uses the simple-counting method by keyword frequency, so it cannot take into account the sematic relationship with the keywords and sematic information such as, how the technologies are used in their technology area and how the technologies affect the other technologies. To automatically analyze unstructured technological information in patents to extract the semantic information, it should be transformed into an abstracted form that includes the technological key concepts. Specific sentence structure 'SAO' (subject, action, object) is newly emerged by representing 'key concepts' and can be extracted by NLP (Natural language processor). An SAO structure can be organized in a problem-solution format if the action-object (AO) states that the problem and subject (S) form the solution. In this paper, we propose the new methodology that can extract the SAO structure through technical elements extracting rules. Although sentence structures in the patents text have a unique format, prior studies have depended on general NLP (Natural language processor) applied to the common documents such as newspaper, research paper, and twitter mentions, so it cannot take into account the specific sentence structure types of the patent documents. To overcome this limitation, we identified a unique form of the patent sentences and defined the SAO structures in the patents text data. There are four types of technical elements that consist of technology adoption purpose, application area, tool for technology, and technical components. These four types of sentence structures from patents have their own specific word structure by location or sequence of the part of speech at each sentence. Finally, we developed algorithms for extracting SAOs and this result offer insight for the technology innovation process by providing different perspectives of technology.

**Keywords :** NLP, patent analysis, SAO, semantic-analysis

**Conference Title :** ICKDML 2016 : International Conference on Knowledge Discovery, Mining and Learning

**Conference Location :** Kuala Lumpur, Malaysia

**Conference Dates :** February 11-12, 2016