Supporting Technical Decision-Making with *InSciTe*®

Seungwoo Lee[1], Mikyoung Lee[1], Hanmin Jung[1], Pyung Kim[1], Taehong Kim[12], Dongmin Seo[1], Won-Kyung Sung[1]

[1]Korea Institute of Science and Technology Information (KISTI),
335 Gwahangno, Yuseong-gu, Daejeon, KOREA 305-806
[2]University of Science and Technology (UST),
113 Gwahangno, Yuseong-gu, Daejeon, KOREA 305-333

{swlee, jerryis, jhm, pyung, kimtaehong, dmseo, wksung}@kisti.re.kr

Abstract. Most researchers are experiencing difficulties in making decisions on their R&D due to information overload. In order to improve research productivity of researchers as well as to support systematic research and development, we have developed InSciTe as a technology intelligence service, which aims to support decision-making processes, especially in establishing R&D strategy. It has been implemented by combining Semantic Web and text mining technologies, analyzes relations among technologies, research agents and research outcomes and provides killer services in the viewpoints of discovery, combination and comparison. Comparison with existing analytic tools will be presented.

Keywords: Decision-making, R&D strategy, technology intelligence, Semantic Web, text mining

1 Introduction

According to a study by Davinci Institute[1], 43% of people are experiencing difficulties in making decisions due to information overload. Researchers in the fields of science and technology desire services that help them to analyze technical data including numerous research outcomes such as papers and patents, and to reflect such data on their research or services so that they can establish an R&D direction for new research domain. In order to improve research productivity of researchers as well as to support systematic research and development, we attempt to develop InSciTe as a technology intelligence service, which supports decision-making processes, especially in establishing R&D strategy.

Technology intelligence refers to activities for supporting an organization's decision-making process by collecting and forwarding information on new technologies [1]. We focus on decision-making researchers who have responsibility for establishing R&D strategy and define technology intelligence service as a provision of insights required for them to establish their R&D strategy or make a decision on their research direction. Our technology intelligence service named

[1] http://www.davinciinstitute.com/page.php?ID=120

InSciTe extracts meaningful technologies and their relations from texts and combines the results with meta-data on a semantic service platform to enhance their analytical values. It analyzes correlations among technologies, research agents and research outcomes, centering on the relations such as competition and cooperation.

There are several representative systems designed to analyze research data, such as VantagePoint[2], Aureka[3], STN AnaVist[4] and Thomson Data Analyzer[5]. VantagePoint and Thomson Data Analyzer are text mining tools for discovering knowledge in structured text datasets. Aureka provides a secure, online environment for searching and storing intellectual property information. STN AnaVist is a powerful interactive analysis and visualization software that offers a variety of ways to analyze search results from scientific literatures and patents. We will discuss and compare our InSciTe with these systems in Section 4.

2 Technologies Used

InSciTe is based on OntoFrame, which is a Semantic Web-based service platform providing implementation infrastructure. OntoFrame aims to search information and discover implicit knowledge in the information for helping users to achieve their needs efficiently [2] [3]. It includes a semantic knowledge management tool named OntoURI, a commercial search engine and a reasoning engine named OntoReasoner. The ontology instances populated by OntoURI are stored and inferred using OntoReasoner, which performs rule-based reasoning based on RDF Semantics and partial OWL Semantics in ways of forward-chaining [4] and also answers to a query represented in SPARQL.

We designed an ontology model to be used for our technology intelligence service. It models research agents such as person (researcher), institution and nation, their research outcomes such as papers and patents, and research topics. It currently consists of 17 classes, 57 datatype properties and 37 object properties. Based on this ontology model, InSciTe currently has 13 million RDF triples which cover 340,000 papers, 70,000 patents, 490,000 researchers, 90,000 institutions and 57,000 research topics. We selected the papers and patents related to green technology domain from NDSL[6] to make focused service possible with not-enough initial data. We will continue to add more papers and patents to augment the quality of technology intelligence service.

InSciTe also adopts SINDI [5] and DISA [6] to extract named entities and their relations. SINDI recognizes core and significant entities from large-scale documents on science and technology domains and generates higher-dimensional, technical knowledge by extracting and processing correlations among the recognized entities. DISA is a sentiment analyzer that extracts opinions using lexico-semantic pattern

[2] http://www.thevantagepoint.com/vantagepoint.cfm
[3] http://scientific.thomson.com/products/aureka/
[4] http://www.cas.org/products/anavist/index.html
[5] http://thomsonreuters.com/products_services/legal/legal_products/intellectual_property/Thomson_Data_Analyzer, which is powered by VantagePoint.
[6] http://www.ndsl.kr/index.do

matching approach. Not only extracting opinion holders, DISA also finds triple relation information.

By using SINDI and DISA, we managed to extract names of technologies and their relations from 410,000 papers and patents in the field of green technology. We have also extracted various relations among the technologies such as element technology, similar technology, competing technology, etc. from PubMed data and Wikipedia as well as the papers and patents. PubMed data and Wikipedia were used to augment the relations extracted from the papers and patents since the number of used papers and patents was not enough to extract rich entity relations. These extracted technologies and their relations are compiled into OntoFrame in RDF format and further utilized for various technology-related services such as technology network and technology-agent map, which will be explained in the following section.

3 InSciTe as a Technology Intelligence Service

InSciTe is designed to help users to analyze patents and papers and to establish their R&D strategy. To differentiate it from existing analytical tools above mentioned, it was devised with the following goals in mind:

- Building fast and automated knowledge;
- Developing effective services conducive to making decisions and;
- Enhancing information accessibility using Linked Data[7].

Main features of InSciTe are as follows. First, it has a through process of Extract-Transform-Load (ETL) and analysis by combining text mining and Semantic Web technologies. It extracts relations between entities using text mining, converts them into semantic data along with meta-data, and then compiles the data on the semantic platform in the RDF triple format for analysis. Second, InSciTe offers enhanced information accessibility through connection with Semantic Web-based open sources represented by Linked Data which are being actively implemented by the U.S., the U.K, and governments in Europe and Oceania. We currently connect to DBpedia[8] and OpenCalais[9] to get descriptions about technology terms and institutions for the reporting service. Third, InSciTe can present multi-faceted viewpoints such as academic and business views by blending heterogeneous sources such as papers and patents. Specifically, users can control the viewpoint dynamically by controlling the ratio between papers and patents used for analysis. Fourth, InSciTe can provide reporting services which summarizes research behavior of institute and research trends related to the given technology or research agent.

InSciTe has four types of key services such as technology-agent composite service, technology-centric service, agent-centric service and reporting service. Technology-Agent Map service, shown in Fig. 1, is one of the representative composite services, which allows users to compare research outcomes and competitive or cooperative relations between research agents – researchers, institutions and nations – for the technologies searched and expanded by the users. From this service, users can also

[7] http://linkeddata.org/
[8] http://dbpedia.org
[9] http://www.opencalais.com/

understand which relevant technologies the major research agents of the concerned technologies are also focusing on.

Fig. 1. Technology-agent (esp., nation) map for keyword 'global warming': the size of circle indicates the strength of correlation and the colors in a row indicates competitive (different colors) or cooperative (same colors) relation between nations

Technology-centric services include technology browser, technology performance graph and irrupted technology-agent trend. Technology browser visualizes the elementary, similar and competitive relationships between technologies (Fig. 2 (a)). Technology performance graph allows users to discover a given technology's current level of maturity (Fig. 2 (b)). Irrupted technology-agent trend service shows entry timings of given technologies and agents (Fig. 2 (c)).

Meanwhile, agent-centric services (Fig. 3) are designed to help users to figure out the competitive/cooperative relations between research agents for a given technology by grouping mutually-cooperating research agents. These services also allow users to understand market trends of a technology. For instance, users can find out whether the market for a technology is led by a sole, representative agent or is being developed through mutual, balanced cooperation between several agents.

Finally, reporting service provides summarization on research trends related to the concerned technologies or research agents. This service was designed to support users' convenience in utilizing the analyzed information.

4 Discussion

There are several existing patent or paper map tools like VantagePoint, Aureka, STN AnaVist and Tomson Data Analyzer that provide similar functions and services to InSciTe. These tools provide users with various complicated quantitative-level

analyses for the data they input. The analyzed results are usually presented in two-dimensional matrices, which can be controlled flexibly by users. They also provide various useful visualization components such as contour map, especially suitable for quantitative-level analysis.

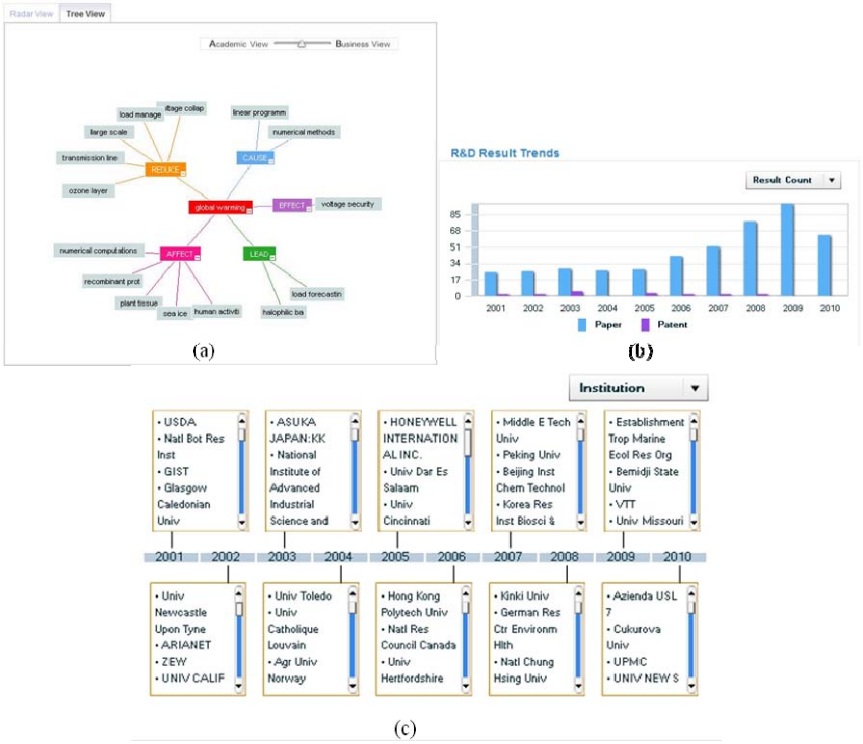

Fig. 2. Technology-centric services for keyword ‘global warming’: (a) technology browser, (b) technology performance graph, and (c) irrupted technology-agent (institution) trend

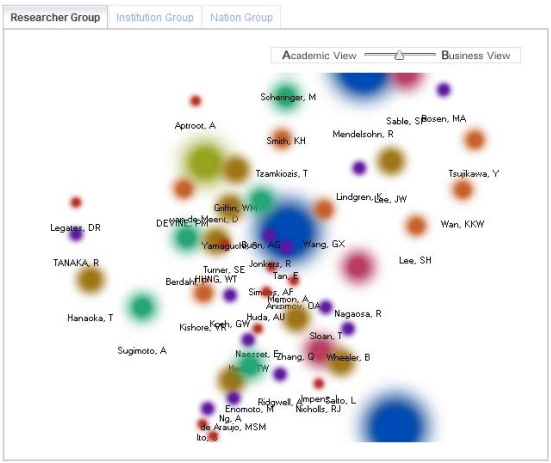

Fig. 3. Researcher groups for keyword ‘global warming’ as one of agent-centric services

However, those tools require users to secure papers and patents in advance and perform pre-processes to make them in system-specific format before importing it into the tools. They also require complicated and skilled techniques for users to get useful analyzed results from the imported data. Most of the tools provide useful quantitative-level analyses such as global trends, each trend by nation/researcher/technology and research network but do not present comparative analyses such as competitive or cooperative relations between research agents in the multi-faceted viewpoints. That is, they do support analyses only in single viewpoint of year, research agent and technology. Some of them even have a limit in the size of data to be processed. VantagePoint, of course, can provide some analysis results in two-dimensional viewpoint by users' selection but many parts of multi-faceted analyses remain to users' work yet. Complicated usage of those tools makes them more appropriate to skilled analysts than general users.

On the contrary, InSciTe was designed to provide diverse analyses in the multi-faceted viewpoint by mutually combining several entities based on possible their relations although it does not provide highly-complicated quantitative-level analysis algorithms and flexibilities in combining the analysis conditions that users want. That is, InSciTe can present the analyzed results in academic, business or mixed viewpoints by blending heterogeneous sources such as papers and patents. Users can control the viewpoint dynamically by controlling the ratio between papers and patents used for analysis. We especially designed and developed InScite to be actively used for establishing R&D strategy, instead of focusing on measuring services. We also designed carefully the easy user interfaces of InSciTe such as one-click services for even general users to have no difficulties in the use. InSciTe internally processes paper and patent data in RDF format. This covers bibliographic meta-data as well as data – named entities and their relations – mined from unstructured text of the papers and patents. It also accesses to external semantic data like DBPedia and OpenCalais to augment the quality of the services, for example, to get descriptions about technology terms and institutions.

To summarize, the comparison of InSciTe to VantagePoint, the most popular one, is given in Table 1.

Table 1. The comparison of InSciTe to VantagePoint

	InSciTe	**VantagePoint**
Data size	~ tens of millions records	~ 20,000 records
Target users	Planner, expert, chief officer, …	Analyst, consultant
DB	Metadata, full-text (DB2OWL)	Bibliographic database (import filter)
Dimension of analysis	Multi-dimensional	2-dimensional (co-occurrence matrices, maps and networks)
Text mining level	Entity/relation extraction	Keyword extraction
Service type/method	Canned services Pull and push services	DIY, scripting Pull services
Others	Ontology model	Expectancy value using Bernoulli process

5 Conclusion

We have developed a technology intelligence service named InSciTe into which Semantic Web and text mining technologies are combined. InSciTe aims to provide insights required for making a decision in the area of R&D by analyzing relations among technologies, research agents, and research outcomes and providing services such as trend, competition and comparison. When compared with existing analytic tools, InSciTe enables multi-dimensional convergence among entities and multi-faceted analysis using Semantic Web technology, while also offering various advantages including extraction of significant entities through text mining technology.

In the future, we plan to introduce additional combined services in the viewpoint of technologies and agents, including prediction, and further to generate institute summary report of higher quality automatically.

References

1. Letizia Mortara, Clive Kerr, David Probert, Robert Phaal: Technology Intelligence-Identifying threats and opportunities from new technologies, University of Cambridge Institute for Manufacturing, ISBN-978-1902546513, 2007.
2. Use Case: OntoFrame 2008 – A Semantic Portal Service of Academic Research Information, 2009. http://www.w3.org/2001/sw/sweo/public/UseCases/OntoFrame/
3. Case Study: Integrated, Connected Search Service for Technical Standards Information, 2010. http://www.w3.org/2001/sw/sweo/public/UseCases/Kisti/
4. Seungwoo Lee, Mikyoung Lee, Pyung Kim, Hanmin Jung, Won-Kyung Sung: OntoFrame S3: Semantic Web-Based Academic Research Information Portal Service Empowered by STAR-WIN, Part II, LNCS6089, 2010.
5. Chang-Hoo Jeong, Sung-Pil Choil, Yun-Soo Choi: Introduction of the Scientific Intelligence Discovery Framework using Grid Computing, International Conference on Convergence Content, 2009.
6. Kyungsun Kim, Jinwoo Park and Junhyung Park: DISA: A Sentiment Analyzer System using Linguistic Pattern Matching Approaches, In Proceedings of ASWC2008, 2008.

Appendix

A. Minimal requirements

- InSciTe is an application that mainly targets researchers who have responsibility for establishing R&D strategy rather than for surveying R&D information. It provides various combined and analyzed results on correlations among technologies, research agents, and research outcomes, centering on their competitive and cooperative relations.
- The data sources cover technical documents including papers and patents. We currently selected 340,000 papers and 70,000 patents, which will be expanded up to hundreds of thousands, related to green technology domain to make focused

services possible with the initial data. The papers came from 109 journals listed in Web of Science and the patents were selected from Korea, US, Europe, Japan and International patents. We also used PubMed and Wikipedia data to augment the relations extracted from the papers and patents and DBpedia and OpenCalais to get descriptions about technology terms and institutions for the reporting service.

- The data used in InSciTe are represented in RDF format. All entities such as technology terms, research agents and research outcomes are represented in URIs and their relations are represented in triples. Such entities and their relations are further processed to derive useful implicit knowledge. The applied Semantic Web technologies give us sufficient flexibility in composing diverse technology intelligence services, which is the main advantage of InSciTe compared to existing analytic tools.

B. Additional features

- InSciTe was designed to be a simple and easy Web application considering researchers unskilled on R&D establishment. The analyzed results are visualized graphically to help users' understand.
- InSciTe currently has about 13 million RDF triples which cover 340,000 papers, 70,000 patents, 490,000 researchers, 90,000 institutions and 57,000 technologies. We continue to add more papers and patents from various journals and international patents to augment the quality of technology intelligence services.
- The advantage of InSciTe is the flexibility that enables to easily compose and extend multi-faceted services for technology intelligence using Semantic Web technologies. For example, it is easy to extend competitive or cooperative groups of researchers into competitive or cooperative groups of institutions or nations based on semantic relations among researchers, institutions and nations.
- There are several analytic tools to be used for supporting technology intelligence, but they mainly consider to be used for highly-skilled analysts and have complicate interfaces difficult to be learnt. To the best of our knowledge, InSciTe is the first intelligence service for unskilled researchers based on Semantic Web technologies.
- In InSciTe, all entities are represented under URI scheme and linked each other based on their possible relations. InSciTe goes beyond pure information retrieval since its retrieval is acted by semantic relations between entities.
- InSciTe provides analyzed results in multi-faceted viewpoints between academic and business. It uses two types of source data such as papers and patents. Papers can provide information in academic viewpoint whereas patents usually provide information in business viewpoint. By fusing these two types of sources, InSciTe can dynamically make up analyzed results in multiple viewpoints.