TechWatchTool: Innovation and Trend Monitoring

Hong Li **Feiyu Xu** **Hans Uszkoreit**

German Research Center for Artificial Intelligence (DFKI), LT-Lab
Alt-Moabit 91c, 10559 Berlin, Germany
{lihong,feiyu,uszkoreit}@dfki.de
http://www.dfki.de/lt/

Abstract

In this paper we present an information service system that allows users to search for the key players of requested technology areas and for their collaboration networks. This system utilizes information extraction and wrapper technologies for detecting persons, organizations, publications and patents as well as relationships among them. Furthermore, it applies relation extraction to detect statements on the web that indicate innovation trends. Various visualization methods are provided to let users monitor key players, their networks and technology trends in a comfortable way.

1 Introduction

The innovation cycle of technologies is getting shorter and shorter. In recent years, many companies became aware of the potential of advanced information technologies for the efficient discovery and analysis of useful information in large volumes of online data such as business news, business reports, scientific publications and patents. Exploring patents or publications is an important approach to analyzing the trends of technology development. Therefore, several systems emerged recently, which attempt to describe and predict the technology development trend based on the analysis of patents or publications (e.g., (Yoon and Park, 2004), illumin8 system[1], Google Trends[2](Rech, 2007), BlogPulse[3] (Glance et al., 2004) and Collexis[4]). Most available systems are mainly based on a combination of statistical methods and string match. There is still a big potential to apply language technologies to this task.

In this paper, we present a system named TECHWATCHTOOL[5], that has already been successfully tested by corporate users. In daily operation, it now aids companies and analysts in detecting emergent technologies and in identifying associated key players, their cooperative networks and new trends that are relevant for their business sector. TECHWATCHTOOL applies methods from bibliometrics, information wrapping, information extraction and data mining. Language technology plays a central role in the extraction of names and technologies. The system monitors technologies with three modules: 1) a retrieval and extraction module for publications and patents for identification of key players and their relations, 2) a trend identification module and 3) an ontology-based navigation module. Furthermore, TECHWATCHTOOL provides different views of the discovered data, which facilitate understanding and interpretation of the results.

The remainder of the paper is organized as follows: Section 2 explains existing systems for technology and trend monitoring. Section 3 introduces the NLP tools used in TECHWATCHTOOL. Section 4 describes our system architecture and the core modules. Section 5 explains the result visualization and presentation. Finally, Section 6 gives a short conclusion.

2 Related Work

Yoon and Park (2004) present a method to create patent networks with text mining methods to investigate the technology development. Patents are represented as nodes in a graph. Similar patents are connected by edges, which are computed automatically from relevant keywords. The system illumin8 implements a semantic search in patent- and web-documents. For a given keyword, the corresponding ontology concepts are identified in the

[1] http://www.illumin8.com/
[2] http://www.google.de/trends
[3] http://www.blogpulse.com/
[4] http://www.collexis.com/products
[5] http://th-ordo.dfki.de/TechWatch_Smila/login.jsp

Proceedings of Recent Advances in Natural Language Processing, pages 660–665,
Hissar, Bulgaria, 12-14 September 2011.

documents. The system provides a modeling of various concepts (e.g. products), but the collaboration networks among the concepts are missing. In addition, illumin8 also illustrates the change of the numbers of active persons or organizations in a certain period. Google Trends is not more than a statistical summarization of its search function (Rech, 2007). As a more advanced example, BlogPulse (Glance et al., 2004) is a system for automatic discovery of trends in blogs. It can find new trends as well as visualize the chronological development of specific terms. BlogPulse extracted trends based not only on terms, but also on videos, news and links which are the targets of daily interests. The system Collexis can discover relationships between elements from different content sources. It can aggregate information from multiple content sources and help to discover potential new hypotheses on large amounts of unstructured contents. All these systems rely more and less on information retrieval technologies and are limited in extracting structured information from free texts.

3 NLP Tools

In TECHWATCHTOOL, named entity (NE) recognition and information extraction (IE) tools are applied to extract named entities (persons, organizations, etc.) and to detect relations or mentions of trends. Two tools are integrated in our system:

1. SProUT as named entity recognizer (Drozdzynski et al., 2004) and
2. DARE as relation extractor and trend sentence detector (Xu et al., 2007; Xu, 2007).

3.1 SProUT

SProUT[6] (Shallow Processing with Unification and Typed Feature Structures) is a platform for development of multilingual shallow text processing and information extraction systems. It is a generic rule-based recognizer to extract named entities or concept terms. Users can write corresponding recognition patterns and specify linguistic resources, such as lexicons, gazetteers and tokenizers. The platform provides linguistic processing resources for several languages including English, German, etc. SProUT uses typed feature structures (TFS) as a uniform data structure for representing the input resources and the recognized named entities. In TECHWATCHTOOL, SProUT is utilized to extract named entities (e.g., persons, organizations and journals) from free texts and to deal with name variants. A special heuristics is implemented in our system via the unification method provided by SProUT, in order to find the equivalent classes of persons and organizations. For example if "Eckhard Beyer" and "Prof. E. Beyer" are the authors of publications about the same technology, they might be identified as name variants of the same person by our method.

3.2 DARE

DARE[7] (Domain Adaptive Relation Extraction) is a minimally supervised machine learning framework for extracting relations of various complexity. It consists two major parts: 1) rule learning, 2) relation extraction. Rule learning and relation extraction feed each other in a bootstrapping framework. The bootstrapping starts from so-called "semantic seed" as a search query, which is a small set of instances of the target relation. (Uszkoreit, 2011) and (Li et al., 2011) describe the application and evaluation of DARE on different corpora for different relation extraction tasks. Currently DARE provides linguistic components which process English and German free texts. In TECHWATCHTOOL, DARE is used to learn linguistic patterns to recognize sentences that potentially contain the trend information and also relations between persons and organizations. To learn patterns from trend sentences, we used the corpus offered by the project partner ThyssenKrupp AG, which is annotated with trend sentences and terms by the experts. From the annotation, we acquire examples as seed for DARE to learn patterns, e.g.,

- ("lithium-ion battery", "car", "future")
- ("Gary Mepsted", "lithium-ion battery"')

The following is an example of trend-statement with its pattern:

pattern: *"power:Verb" ([subj:Noun], [obj:"car"], [mod:"future"])*

trend-statement: Lithium batteries power hybrid cars of future[8]

[6] http://sprout.dfki.de/index.html

[7] http://dare.dfki.de

[8] http://www.reuters.com/article/2007/06/21/environment-batteries-lithium-saft-dc-idUSL2055095620070621

To learn patterns for recognizing the relation between persons and their positions in an organization, we use the Penn Treebank as our linguistically annotated corpus and some examples of the following triple:

<person, organization, position>

as start seeds.

4 System Architecture

TECHWATCHTOOL is a web application for multiple users, implemented in Java6. It has three modules dealing with different scenarios:

1. Searching and identification of key players and their collaboration network from patents and publications
2. Identification of trends for an area
3. Ontology-based navigation of a specific domain

4.1 Search and Identification of Key Players

Scientific publications and patents are two important indicators of technology development. Authors, applicants or owners of these two resources are active persons or organizations in their respective areas. Our task is to extract these active persons and institutions, identify their relationships and discover key players among them.

Fig. 1 shows the workflow and components of this module.

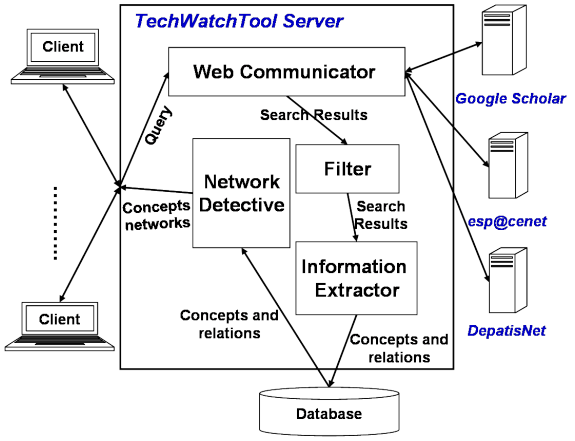


Figure 1: Workflow of search and identification of key players

Given a user query, for example, a technology term (e.g., "laser beam welding") or a company name (e.g. "NISSAN Motor"), the *Web Communicator* will acquire the relevant publications and patents from three resources: Google Scholar[9], esp@cenet[10] and DepatisNet[11]. Three wrappers are implemented to extract relevant concepts such as publication names, publication types, patent names, applicants, owners and author names and their relations by utilizing the named entity recognition tool SProUT.

The ranking of a key player is based on the number of publications or patents published or owned by a person or an organization, the recency of the publications and patents and the connectivity of the person and the organization in their technology community.

$$score(p \in P) = \frac{|P|}{index\ of\ p\ in\ P} \quad (1)$$

where P is the search result list of patents or publications from the three web resources.

$$score(t) = \begin{array}{c} \alpha \times \sum_{pat \in Pat(t)} score(pat) + \\ (1-\alpha) \times \sum_{pub \in Pub(t)} score(pub) \end{array} \quad (2)$$

where t is a player that can be either a person or organization, $Pat(t)$ is the patent set belongs to this player as the inventor or owner and $Pub(t)$ is the corresponding publication set. α is the scoring parameter ranged from 0 to 1. The default value is set to 0.5.

The identified key players' names can be used as new search queries to search for new patents and publications about relevant technologies.

4.2 Identification of Technology Trends

Fig. 2 shows the detailed workflow of this module. The task of technology trend identification is to extract statements indicating the future trends of a specific technology expressed by key players. TECHWATCHTOOL retrieves firstly relevant documents with the Google Custom Search Engines[12], which are defined by the experts of the user company. Linguistic patterns are applied to the documents to recognize sentences that potentially contain the trend information. The linguistic patterns are determined in two ways:

[9] http://scholar.google.de/, a search engine for scientific publications

[10] http://ep.espacenet.com/, the European patent web server

[11] http://depatisnet.dpma.de/DepatisNet/, the German patent web server

[12] http://www.google.com/cse/

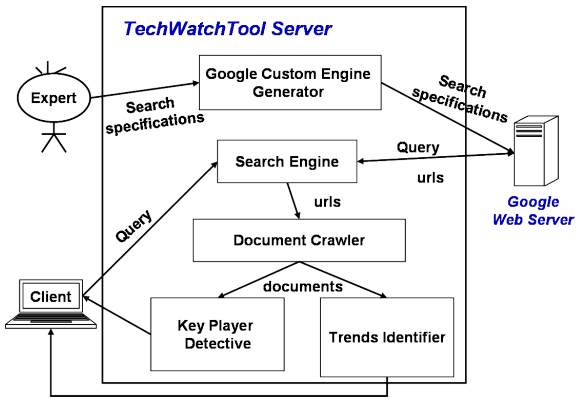


Figure 2: Trend identification

1. the linguistic experts define and evaluate the initial set of patterns in the form of regular expressions;

2. the machine learning system DARE (see Section 3.2) acquires additional patterns by learning rules from the dependency structures.

The regular expressions are designed based on the lexical indicators of the potential trend statements. The domain experts highlight the texts as samples for designing and scoring the patterns. Text statements that match these patterns are considered the indicators of potential trends. In the following, we show an example of the trend patterns and a statement matching them:

pattern1: *future of (.){0,20}car*

pattern2: *in (the)? future*

trend-statement: Mass.-based A123 Systems is now worth nearly $2 billionindicating huge investor confidence **in the future of electric cars**, plug-in hybrids, and the batteries that make them go.[13]

As described in Section 3.2, we use the DARE system to identify the text statements and trend terms. Compared to the regular expression-based patterns, the rules learned by DARE are more accurate because they consider the syntactic structures and more bigger linguistic contexts. Therefore, the recognition is more precise. Furthermore, DARE is able to correct and update the rules when more queries and more documents are generated through the users. On the other hand, the dependency structures in DARE system are fairly strict, therefore, not as robust as the regular expressions. Therefore, we combine both methods to detect more trend statements without compromising on the accuracy.

Using this module, TECHWATCHTOOL can also identify the key players who are active in a certain domains without identified connections to any publications or patents. Such key players may be large corporations, department leaders or managers. The persons and organizations are evaluated based on their relevance to the given query

$$score(t) = \frac{occurrences\ of\ t\ with\ the\ query\ in\ sentence}{occurrences\ of\ t\ in\ document} \quad (3)$$

The relations between these persons and organizations are detected by patterns acquired by DARE as described in Section 3.2. The following is an example sentence for the given query *machine learning*:

> *One of those bright-eyed children was **Christopher Bishop**, now a partner at **Microsoft Research** in Cambridge and **a leading expert** in machine learning.*[14]

This module can be connected with the patent and publication search module to find out whether the identified key players are also owners of any publications or patents.

4.3 Interactive Ontology-based Navigation

TECHWATCHTOOL allows users of a specific technology domain to monitor the technology development via a web-based ontology-based navigation user interface. An ontology for a specific technology domain is usually provided by the experts of the user companies. Users can zoom into the ontology and find concepts (named by technology terms) and their subconcepts and obtain information about selected items. The information can contain a description of this concept, recent publications and patents, its new key players and new trends in the area. Fig. 3 displays a screen shot of the web interface.

5 Data Visualization and Result Presentation

It is always a challenge for web applications to present users the results in an intuitive way (Andrews, 1995; Rohrer and Swing, 1997). TECHWATCHTOOL allows users to have at least three

[13] http://www.hybridcars.com/news/investors-embrace-a123-lithium-new-ethanol-26126.html

[14] http://www.theengineer.co.uk/in-depth/interviews/machine-learning-expert-prof-chris-bishop/1008899.article

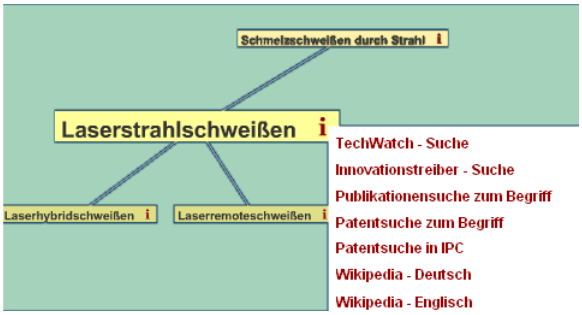
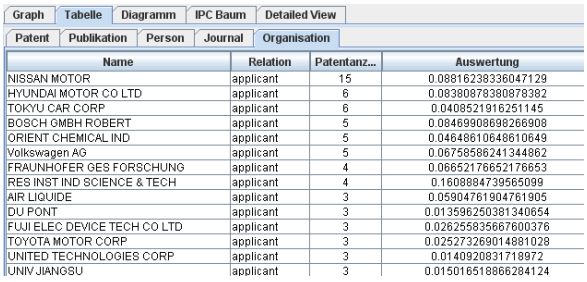


Figure 3: Ontology-based navigation



Name	Relation	Patentanz...	Auswertung
NISSAN MOTOR	applicant	15	0.0881623833604712 9
HYUNDAI MOTOR CO LTD	applicant	6	0.0838087838087838 2
TOKYU CAR CORP	applicant	6	0.0408521916251145
BOSCH GMBH ROBERT	applicant	5	0.0846990869826690 8
ORIENT CHEMICAL IND	applicant	5	0.0464861064861064 9
Volkswagen AG	applicant	5	0.0675858624134486 2
FRAUNHOFER GES FORSCHUNG	applicant	4	0.0665217665217665 3
RES INST IND SCIENCE & TECH	applicant	4	0.1608884739565099
AIR LIQUIDE	applicant	3	0.0590476190476190 5
DU PONT	applicant	3	0.0135962503813406 54
FUJI ELEC DEVICE TECH CO LTD	applicant	3	0.0262558356676003 76
TOYOTA MOTOR CORP	applicant	3	0.0252732690148810 28
UNITED TECHNOLOGIES CORP	applicant	3	0.0140920831718972
UNIV JIANGSU	applicant	3	0.0150165188662841 24

Figure 5: Table view: sortable table

views onto the results: i) graph view ii) table view iii) diagram and chart view.

The graph view is suited for presenting the collaboration networks among active persons and organizations and their relations to publications and patents. Fig. 4 shows an example of such networks. The advantage of the graph viewer is that

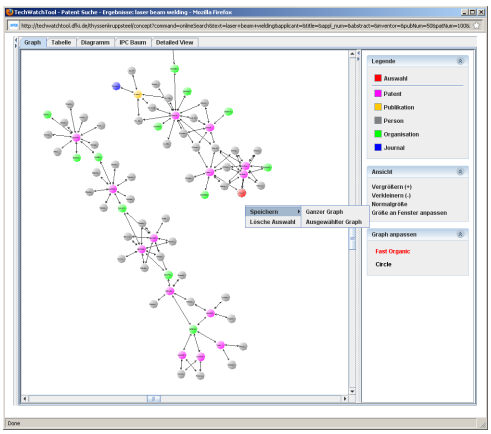


Figure 4: Interactive graph viewer for collaboration networks among persons, organizations, journals, publications and patents

users can monitor and observe new clusters of collaborations in a very straightforward way. Furthermore, the viewer offers to the users various convenient functionalities such as graph layouts, zooming and focusing functions as well as customization of the color scheme.

In order to assist users in finding key players quickly, the table view provides sortable tables containing relevance ranking information about persons or organizations.

For monitoring the technology development in a certain time interval, a diagram viewer is included in TECHWATCHTOOL as depicted in Fig. 6. This diagram viewer provides the total number of publications or patents for each year within the time interval and offers users a fairly direct overview of the historic development of the technology. Furthermore, users can also compare the changing proportions between publications and patents (Fig. 7).

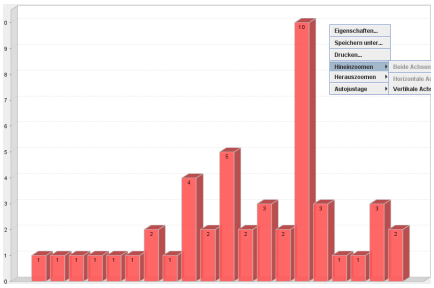


Figure 6: Diagram viewer for publications and patents

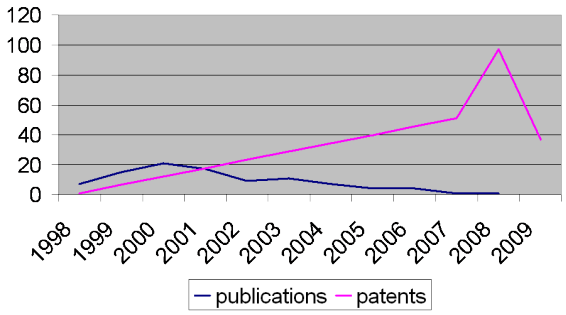


Figure 7: Relationships between publications and patents

Furthermore, TECHWATCHTOOL features many other visualization functions such as tree viewer, html viewer etc. All these visualization tools can export their graphs or tables into files when required. The export function facilitates the further processing of the result information.

6 Conclusion and Future Work

This paper describes and demonstrates a provenly useful application that assists experts in monitoring new technology developments and detecting new technology trends. The system combines information wrapping, information extraction and data mining technologies and provides different views of result presentation. Through these means, users can access and interpret the information in a very convenient way and thus gain valuable new insights.

As described in Section 3, the recognition of concept terms relies on the NLP tools. Therefore, the errors of NLP tools can damage the accuracy of TECHWATCHTOOL analysis. Meanwhile the patent and publication analysis is based on the search results of the web search engines that can neither guarantee precision nor recall. Therefore, avoiding the negative consequences of these factors and evaluating the quality of the TECHWATCHTOOL system proper remains an open challenge. It is also very difficult to automatically assess the extraction and identification results of the trend search module. We plan to evaluate it manually by annotating a small sample of documents. The identification algorithm of the trend search module still needs to be improved. We plan to run the DARE rule-learning system during the application of TECHWATCHTOOL automatically to acquire new patterns and to validate the learned patterns. We also intend to update the ontology by the new technology terms learned from document via the trend search module. Our current method for evaluating the persons and organizations in the trend module still produces errors. It happens that unrelated persons or organizations occasionally occur together with the given query pattern. This over-detection will hopefully be alleviated by NLP tools that utilize the syntactic structures of the sentences, such as DARE does.

Acknowledgments

The research reported in this paper was initialized in the context of industrial projects funded by ThyssenKrupp AG and was further developed in the project Theseus Ordo (funded by the German Federal Ministry of Economics and Technology (BMWi) through the contract 01MQ07016). Many thanks to Peter Seyfried, Ralf Sünkel and Haydar Mecit of ThyssenKrupp for their valuable suggestions, comments and ccoperation.

References

K. Andrews. 1995. Visualizing cyberspace: Information visualization in the Harmony Internet browser. *Proceedings of Information Visualization*, pages 97–104.

Witold Drozdzynski, Hans-Ulrich Krieger, Jakub Piskorski, Ulrich Schäfer, and Feiyu Xu. 2004. Shallow processing with unification and typed feature structures — foundations and applications. *Künstliche Intelligenz*, 1:17–23.

N. Glance, M. Hurst, and T. Tomokiyo. 2004. Blogpulse: Automated trend discovery for weblogs. In *WWW 2004 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics*, volume 2004. Citeseer.

Hong Li, Feiyu Xu, and Hans Uszkoreit. 2011. Minimally supervised rule learning for the extraction of biographic information from various social domains. In *Proceedings of RANLP 2011*.

J. Rech. 2007. Discovering trends in software engineering with google trend. *ACM SIGSOFT Software Engineering Notes*, 32(2):1–2.

R.M. Rohrer and E. Swing. 1997. Web-based information visualization. *IEEE Computer Graphics and Applications*, 17(4):52–59.

H. Uszkoreit. 2011. Learning relation extraction grammars with minimal human intervention: strategy, results, insights and plans. *Computational Linguistics and Intelligent Text Processing*, pages 106–126.

Feiyu Xu, Hans Uszkoreit, and Hong Li. 2007. A seed-driven bottom-up machine learning framework for extracting relations of various complexity. In *Proceedings of ACL 2007, 45th Annual Meeting of the Association for Computational Linguistics*, Prague, Czech Republic, 6.

Feiyu Xu. 2007. *Bootstrapping Relation Extraction from Semantic Seeds*. Phd-thesis, Saarland University.

B. Yoon and Y. Park. 2004. A text-mining-based patent network: Analytical tool for high-technology trend. *The Journal of High Technology Management Research*, 15(1):37–50.