Research and Practice on the Construction Plan for a Thematic Patent Early Warning Platform (Postprint)

Authors: Wang Li, Ding Yingjie, Wu Ming

Date: 2017-11-08T00:00:00+00:00

Abstract

【Purpose】To develop a construction plan for a thematic patent early warning platform, providing a solution for long-term thematic tracking and early warning analysis, thematic data reuse, and other related work. 【Method】The platform integrates open-source code platforms and tools (DSpace, OpenRefine, ECharts, VOSviewer, etc.) to implement functions such as storage, tracking, classification, cleaning, analysis, and management of thematic data. 【Results】The extreme ultraviolet lithography technology theme was selected for application practice, testing and resolving detailed issues during the implementation process, and verifying the feasibility and effectiveness of the thematic patent early warning platform. 【Limitations】The current thematic patent early warning platform requires further optimization in aspects such as full automation of data processing, standardization of data analysis metrics, and implementation of correlation in content mining. 【Conclusion】The functions implemented by the thematic patent early warning platform hold practical significance for timely tracking, early warning, and classified management of technology patents throughout the technology R&D lifecycle.

Full Text

Preamble

Title: Research and Practice on the Construction Scheme of a Subject-Based Patent Early Warning Platform

Authors: Wang Li[1],[2], Ding Yingjie[1], Wu Ming[1]
[1](National Science Library, Chinese Academy of Sciences, Beijing 100190, China)
[2](University of Chinese Academy of Sciences, Beijing 100049, China)

Abstract: [Objective] This paper proposes a construction scheme for a subject-based patent early warning platform, providing a solution for long-term subject tracking, early warning analysis, and subject data reuse. [Methods] The platform integrates open-source platforms and tools (including DSpace, OpenRefine, ECharts, VOSviewer, etc.) to implement functions such as storage, tracking, classification, cleaning, analysis, and management of subject data. [Results] The extreme ultraviolet lithography technology subject was selected for application practice to test and resolve detailed issues in the implementation process, thereby verifying the feasibility and effectiveness of the subject-based patent early warning platform. [Limitations] The current platform requires further optimization in terms of fully automated data processing, indicator-based data analysis, and implementation of content mining associations. [Conclusions] The functions implemented by the subject-based patent early warning platform have practical significance for timely tracking and early warning of technology patents and their classified management throughout the technology R&D lifecycle.

Keywords: Subject-based system; Patent early warning; System construction

Classification Numbers: TP392; G353

Patent information plays a crucial supporting role in economic and social development as well as enterprise innovation activities. Patent early warning involves retrieving and analyzing patent information to research and predict potential patent risks, thereby supporting response strategies [1]. The stakeholders involved in patent early warning range from regional and national entities to R&D teams. Throughout the technology innovation and industrial development lifecycle, these stakeholders need to capture relevant patent information in a timely and sensitive manner, analyze and predict patent risks, and formulate appropriate responses.

Since patent risks throughout the technology innovation and industrial development lifecycle are dynamically changing, patent early warning work for specific innovation subjects is time-sensitive. From patent retrieval to analysis, and from risk identification to assessment, manual patent early warning cannot monitor patent information in real time, and process patent data is often used only once. These drawbacks hinder the timeliness of patent early warning and the reuse of patent data. As the primary carrier of patent information, the effective utilization of patent data is crucial in patent early warning work. In today' s digital era, where the network has increasingly become the most important channel for scientific and technological exchange and dissemination, using online platforms to track patent information in real time for early warning purposes holds practical significance.

Through our investigation and use of relevant domestic and international systems and platforms, we identified several limitations: (1) Patent analysis platforms can adequately meet single-analysis needs but lack integrated functionality for patent early warning. While some platforms offer partial early warning capabilities, they suffer from inflexible subject creation, low customization, or

inability to track in real time. For instance, the Thomson Innovation platform enables data tracking for specific subjects through alert creation but lacks data cleaning and classification management functions. The Orbit analysis platform allows data cleaning and analysis of fixed patent datasets through workspace folders but cannot perform real-time tracking or classification management. (2) Enterprise competitive intelligence systems generally follow the logical framework of intelligence collection, analysis, and service, with complex data sources dominated by unstructured information. However, they lack specificity for subject-based patent early warning and do not emphasize cleaning of multi-source data. For example, the Goonie competitive intelligence system comprehensively integrates and utilizes various information sources to provide enterprise competitive intelligence, but the information is too extensive for technology R&D innovation purposes. (3) Self-built platforms are developed according to patent management systems, patent intelligence analysis frameworks, or patent information value systems. For example, the Institute of Computing Technology of the Chinese Academy of Sciences developed a patent value analysis and rating electronic system focusing on patent value assessment, inventor self-evaluation, and expert review.

To address these issues, this paper proposes an implementation scheme for a subject-based patent early warning platform. Built according to stakeholders' specific needs for technology innovation and management, the platform effectively integrates the patent intelligence analysis process to enable customized patent classification management and early warning for various segments of technology or industrial chains. The data subjects can range from entire industry patent datasets to specific product or technology patent collections. Functionally, stakeholders can use the platform to form subject-specific patent datasets, classify and manage key R&D technologies, monitor the latest technology development trends, understand competitors' technical capabilities, and achieve patent early warning analysis and data reuse. Technically, the scheme integrates open-source platforms and tools to develop functional modules for storage, tracking, classification, cleaning, analysis, and management of subject patent data, thereby saving system development time and costs.

2. Platform Design and Implementation

Patent data serves as the primary carrier of patent information and constitutes the main analytical object of patent early warning work. The openness of patent data provides a foundation for the collection, organization, and processing capabilities of our patent early warning platform. Patent early warning requires comprehensive and timely tracking of patent information, yet open patent data resources are numerous with non-uniform metadata and content formats across different sources. The subject-based patent early warning platform must enable automatic collection of subject data through customized R&D subject configuration, matching specified content and structurally extracting patent metadata information from various patent information resources while achieving local stor-

age. The platform needs to normalize patent data from different sources through unified customized metadata and perform data deduplication by setting unique patent data identifiers. With data processing functions to unify content formats, the platform helps stakeholders dynamically capture and track patent information while providing high-quality data for early warning analysis. The functional framework of the subject-based patent early warning platform is shown in Figure 1 [Figure 1: see original paper]. Based on this framework, the platform focuses on implementing five core functions: subject customization, data collection, automatic classification, data processing, and data analysis.

(1) Subject customization refers to configuring early warning targets through the development of comprehensive retrieval strategies. Before formulating retrieval strategies, adequate preparatory work is required, including subject domain investigation, research on term diversity, pre-retrieval of target patent resources, and retrieval strategy adjustment and refinement. The completeness of retrieval strategies directly affects the effectiveness of subsequent tracking data. Through subject customization, stakeholders can achieve personalized tracking and early warning for their required subjects.

(2) Data collection involves periodic harvesting of target patent resources by constructing a series of distributable and deployable web-oriented crawlers for precise acquisition [2]. During the data collection phase, metadata items from target patent resources are normalized to achieve unified description of data from different sources, and data deduplication is performed using unique patent data identifiers. Through data collection, stakeholders can achieve real-time tracking of customized R&D subjects.

(3) Automatic classification refers to the customized categorization of collected information. By developing comprehensive matching strategies, the platform enables thematic classification management: each collected patent information item undergoes subject identification and is automatically assigned to the corresponding thematic category while achieving automatic indexing of patent information. Through automatic classification, stakeholders can implement classified tracking and management of customized subjects.

(4) Data processing involves handling collected patent information, including data cleaning and format processing. The diversity of patent information resources leads to inconsistent cataloging formats. While the data collection phase normalizes metadata naming, collected patent information still exhibits cataloging rule diversity. One objective of the data processing phase is to unify these cataloging formats. Uncleaned data commonly suffers from non-standard naming—for example, IBM may appear as IBM, IBM Corp., or International Business Machines Corporation, with subsidiary companies, regional branches, and other legal entities also present. Without standardization through data cleaning, tracking and early warning analysis for applicants would lose accuracy. Therefore, another pur-

pose of the data processing phase is to achieve data cleaning. Stakeholders can ensure the standardization and effectiveness of early warning information by saving data processing rules.

(5) Data analysis provides automatic early warning analysis services for stakeholders. Through the aforementioned series of processes, the data analysis phase delivers services including tracking of key institutions, identification of important inventors, and revelation of hot topics. Simultaneously, stakeholders can perform multi-angle analyses according to their needs to deeply explore early warning information.

3. Key Technical Implementation Methods

Based on the functional framework of the subject-based patent early warning platform, we extended and developed the system using the open-source software DSpace 4.2 [3-4] and integrated OpenRefine, ECharts, and VOSviewer to implement relevant functions. DSpace is a Java-based open-source system with comprehensive metadata definition, local hierarchical data storage, and data indexing and retrieval capabilities, making it the preferred system for developing subject-based patent platforms. Leveraging DSpace' s metadata functionality, the subject-based patent early warning platform standardizes the data fields being monitored. DSpace' s Community and Collection structures provide the technical foundation for the platform' s classification management. To achieve more accurate early warning analysis, the platform requires standardized processing of collected data. OpenRefine [5], with its data processing capabilities and open-source nature, becomes the optimal choice for implementing data processing functions. The platform integrates OpenRefine to realize data cleaning, processing, and preservation of processing rules. While DSpace provides one-dimensional statistical analysis, the platform achieves custom multi-dimensional combined analysis through secondary development, thereby expanding the depth and breadth of early warning capabilities, and integrates ECharts and VOSviewer for analysis visualization. The technical framework of the subject-based patent early warning platform is shown in Figure 2 [Figure 2: see original paper].

The construction of the subject-based patent early warning platform addresses four key technical aspects: customized data collection, automatic classification, data cleaning, and data analysis. The implementation flow of these key technologies is shown in Figure 3 [Figure 3: see original paper].

3.1 Data Collection

To achieve comprehensive monitoring, target patent data may originate from different patent information websites with significantly varying content structures. Additionally, website updates and upgrades can cause changes in page structures, imposing high adaptability requirements on the system. The patent early warning platform must possess flexible analytical capabilities to automat-

ically analyze and collect web content from target patent information resources, storing collected data locally according to pre-defined metadata. The system primarily utilizes Apache' s HttpClient to simulate browser functionality: (1) HttpClient is encapsulated using the singleton pattern, and Double-Check is employed to resolve the issue of creating multiple instances during multi-threaded collection, as shown in the "Initialize Simulated Browser" step in Figure 3. (2) HttpClient simulates browser functionality to achieve collection. Different patent information websites have varying retrieval rules; pre-customized retrieval strategies are configured according to the target website' s rules, and data retrieval is performed based on these configured strategies during collection, similar to the manual retrieval process. (3) The jsoup component is used to extract collected HTML-format data by configuring HTML element selectors to establish one-to-one correspondence with DSpace metadata, thereby storing patent information content locally in the configured metadata format.

3.2 Automatic Classification

To achieve classified management and early warning for various segments of technology or industrial chains and different branches of industries, the automatic classification and early warning function is essential for the patent early warning platform. This study employs classification configuration to implement automatic classification management. Classification configuration information is stored using XML and associated with Communities and Collections in DSpace to achieve classified data storage. First, Communities are used to implement pre-defined classification settings for various segments of technology or industrial chains and industry branches, facilitating front-end display on the patent early warning platform. Second, Collections are used to define access classifications by setting matching rules for each collection, enabling collected data to be classified and stored according to defined rules and thereby achieving automatic classification of patent data. The implementation of automatic classification leverages DSpace' s powerful search functionality: for data matching preset classification rules, DSpace' s built-in application programming interface is used to move the data into the corresponding Collection.

3.3 Data Processing

Despite metadata normalization during the data collection phase, data from various information websites still suffers from inconsistent cataloging formats and non-uniform content expression. Therefore, data processing and cleaning functions are indispensable for the subject-based patent early warning platform. The open-source data processing tool OpenRefine [6] meets these requirements. However, OpenRefine' s data cleaning is not repeatable—a single data processing procedure cannot serve as a template for subsequent processing. Secondary development is required to templatize data cleaning, enabling automatic application of identical processing to updated monitoring data and thereby avoiding repetitive work. OpenRefine is project-based, and its internal processing records

focus on data changes rather than operational steps, making it unsuitable as a universal data processing rule. Through analysis, we discovered that OpenRefine' s data processing operates in a command-based pattern, where each operation is a command. By persistently storing each command operation, we can record OpenRefine' s operational steps and then replay these stored steps on different data, thereby achieving templated data processing using OpenRefine.

3.4 Data Analysis

After collection, classification, and cleaning, patent data is stored in high quality within the patent early warning platform, enabling timely tracking. The platform requires data analysis to deeply explore subject patent information and achieve multi-dimensional early warning effects. While DSpace can perform one-dimensional patent data analysis, such as statistical analysis of institutions, inventors, and time periods, this study conducted secondary development to enhance the platform' s analytical capabilities for multi-dimensional combined analysis. The subject-based patent early warning platform utilizes Solr' s Facet functionality as the foundation for data analysis development. One-dimensional statistical analysis can directly employ Solr' s Facet function, while the system develops metadata display selection functions to enable customized front-end interface displays. To implement customized combined analysis [5], analysis dimensions must first be determined. Second, in system implementation, statistical analysis is performed on the first dimension, followed by combined analysis with the second dimension based on the resulting data subset. For example, when analyzing trends of major institutions, the analysis combination of institution and year must be determined. The analysis requires a TOP query on institutions to extract the subset of top institutions, then uses Facet' s statistical function to analyze their annual data (specifiable as time intervals or discrete values), and finally employs the integrated ECharts [7] for combined analysis visualization. The subject-based patent early warning platform has conducted preliminary exploration of text mining for content, integrating VOSviewer [8] for thematic clustering analysis. However, the current clustering lacks sufficient association with the data, which represents a problem to be addressed in future work.

4. Application Practice

Using the construction practice of the extreme ultraviolet (EUV) lithography technology patent early warning platform as an example, we illustrate the implementation of the proposed patent early warning analysis platform. The EUV lithography technology patent early warning platform can automatically collect, update in real time, clean, and analyze EUV lithography patent data. Additionally, combining the actual requirements of EUV lithography technology, it implements automatic classification functionality, as shown in Figure 4 [Figure 4: see original paper].

The subject-based patent early warning platform establishes unified metadata

rules based on patent data characteristics (applicable to all patent data from different sources). As shown in Figure 5 [Figure 5: see original paper], elements such as name, metadata element, and metadata qualifier are unified for all patent data. Data from different sources requires configuration of targeted HTML element selectors to map them to the platform' s metadata. Figure 5 illustrates the metadata configuration for patent data from freepatentsonline in the EUV lithography technology patent early warning platform, thereby achieving unified platform data structure. The platform uses the patent application number ("Application Number") as the unique identifier for data deduplication. At this stage, the original dataset contained 7,356 records, which was reduced to 5,787 records after system deduplication. The platform monitors and presents updated data in real time through the "Latest Submissions" module on the homepage, achieving early warning tracking.

According to the classification tracking and management requirements of the EUV lithography early warning platform (see Figure 4(a)), corresponding rules are configured for each category, as shown in Figure 6 [Figure 6: see original paper]. This enables automatic classification of patent data, with data unrecognizable by rules supplemented through manual interpretation. The classification management for EUV lithography technology can be navigated from the homepage (see Figure 4(b)).

High-quality data is fundamental to obtaining accurate patent early warning analysis results. After completing EUV lithography technology patent data collection, the platform's data processing function performs data cleaning to obtain unified and standardized data. As shown in Figure 7 [Figure 7: see original paper], the unprocessed EUV lithography technology patent data indicated that Nikon owned 148 patents, which increased to 266 after processing, demonstrating that unprocessed data can mislead patent early warning results. The data processing procedures of this patent early warning platform are automatically saved, as shown in Figure 7, to facilitate subsequent real-time data updates.

With data processing rules established, the EUV lithography technology patent early warning analysis platform can effectively achieve automatic collection, classification management, and real-time updates, and possesses high-quality data for early warning analysis. The platform' s analysis functions enable patent early warning analysis for EUV lithography technology, with results flexibly customizable for display on the platform homepage as navigation modules. The homepage provides early warning analysis navigation modules for patent application dates, applicants, inventors, etc. The platform supports two-dimensional combined analysis, allowing users to freely combine dimensions for analysis and visualization according to specific requirements, as shown in Figure 8 [Figure 8: see original paper].

5. Future Work

The construction scheme of the subject-based patent early warning platform provides a feasible customized solution for subject research, subject management, and subject intelligence work that requires long-term subject tracking, early warning analysis, and subject data reuse. This paper elaborates on the construction scheme and technical implementation of the subject-based patent early warning platform and validates its feasibility and effectiveness through experimental analysis based on the construction process of the "Extreme Ultraviolet Lithography Patent Early Warning Analysis Platform." The developed "Subject-Based Patent Early Warning Platform" still requires improvements and enhancements in several areas, such as: optimization of the underlying platform, full automation of data processing, indicator-based data analysis, and implementation of content mining associations—these will be the focus of future practical work.

References

[1] Zhang Yong. Patent Pre-Waring—from Risk Management to Innovative Competition [M]. The 1st Edition. Beijing: Intellectual Property Publishing House, 2015: 26-28.

[2] Zhang Zhixiong, Zhang Xiaolin, Liu Jianhua, et al. The Ideas and Methods of Structural Monitoring of the Scientific and Technical Information Resources on the Web[J]. Journal of Library Science in China, 2014, 40(4): 4-15.

[3] Zhu Zhongming, Ma Jianxia, Chang Ning, et al. An Implementation of a DSpace-based Disciplinary Repository System [J]. New Technology of Library and Information Service, 2006 (7): 10-14.

[4] DuraSpace. DSpace 4.x Documentation [R/OL]. [2014-08-20]. http://www.yok.gov.tr/documents/7166509/
Manual+4.x.pdf/4ac490ee-9a24-4edd-90b7-a894134c9641.

[5] Wang Li. Solution Research of Open-Source/Fee-Free Tools Used in Full-Process Patent Analysis [J]. Information Studies: Theory & Application, 2016, 39(1): 118-122.

[6] Verborgh R, De Wilde M. Using OpenRefine[M]. Packt Open Source, 2013: 21-64.

[7] ECharts Team. Getting Started [R/OL]. [2016-05-20]. http://echarts.baidu.com/echarts2/doc/doc.html.

[8] Centre for Science and Technology Studies, Leiden University. VOSviewer-Getting Started [OL]. [2015-10-10]. http://www.vosviewer.com/getting-started.

Author Contributions: Wang Li: Conceived the research idea, designed the research plan and technical implementation scheme, conducted experiments, collected, cleaned, and analyzed data, and drafted, revised, and finalized the manuscript; Ding Yingjie: Implemented and optimized the technology, and revised the manuscript; Wu Ming: Conceived the research idea.

Conflict of Interest Statement: All authors declare no conflict of interest.

Supporting Data: Supporting data is self-archived by the authors, E-mail: wangli@mail.las.ac.cn.
[1] Wang Li. CAScode.zip. Original code for the subject-based patent early warning platform.
[3] Wang Li. Dspace_{import}.zip. Data for the extreme ultraviolet lithography technology patent early warning platform.

Received: 2016-06-29 **Revised:** 2016-07-28

Note: Figure translations are in progress. See original paper for figures.