A Knowledge Discovery Framework for XML-Literature-Data

Lixue Zou*, Li Wang, Xiaoli Chen, Xiwen Liu

zoulx@mail.las.ac.cn

National Science Library, Chinese Academy of Sciences
Contents

- Literature Review
- Motivation
- Data Processing Methodology
- Literature Data Mining
  - Concepts and Topic Clustering
  - Substances and roles
  - Link scientific research to industry
- Conclusion
- Future Work
Literature Review

**XML-Literature-Data collection:**
Scientometric analysis, text information extraction and mining have recently been applied to knowledge discovery using literature data modeled using XML, including publications or patent data. The existing methods proposed some methods using either the paper data or the patent data from XML-data. However, the method of extracting the publications and patent data from the same XML-data file has not been seen.

**Different source data mining:**
Scientific research and development play important roles in enhancing national competitiveness, so knowledge discovery of literature data becomes a strategic endeavor. These publications or patents data are retrieved from different databases that do not share the same indexing system, which can not allow us to conduct a comparative study at the same level.
Motivation

- CAplus, a database of Chemical Abstracts Service, which is the world’s largest repository of information on chemistry and related publications, provides the XML-data and covers both papers (Types of publications include journal articles, preprints, conference articles, dissertations, and books) and patents in one database.

- One of the merits is that all the data are provided in the same indexing system, including the concepts, substances and roles, commercial or government entity, source of publication, and various other data entities. Thus, the indexing terms can be used to deep mining and make comparisons between the papers and patents.
Motivation

• Our study focuses on the knowledge discovery framework for XML-Literature-Data based on the CAplus database.

• This study presents the methods of text information extraction and text mining on xml-data from CAplus database.

• The integrative use of indexing data on papers and patents of CAplus and the systematic exploration and comparative study of the distribution trends in topics, substance roles, and industrial mapping are distinctive and insightful.

• This study provides a valuable reference for scientists and developers, policy makers, industry and business.
Data Processing Methodology

Data Collection → Data Exploration → Data Transformation → Run the Java Program to Transform the Data

Knowledge Discovery

Data Collection → Data Exploration → Data Transformation → Data Cleaning → Data Validation And Verification

Featured Data Mining
Data Exploration

- **Structure of the data**
  Documents, Indexing, Family, Substances, Nomenclature, Keymap

- **Attributes collected:**
  - Title
  - Author
  - Abstract
  - Date
  - Concept
  - Substance
Data Transformation

- **Keymap analysis:**
  Analysis of document centric key map between all documents and any associated indexing.

- **Reclassification:**
  Each file contains one document and associated indexing files, and separates the papers from the patents.

- **Data extraction:**
  Extraction for each attribute for papers and patents.

- **Output:**
  Comma-Separated Values

- **Challenge:**
  Substances need to match its function.
Data Cleaning and Statistics

• **Data Cleaning:**
  Remove noise and merge different spelling of one entity.

• **A case study: Global Graphene Research**

• **Data statistics:**
  - 78756 papers, 23057 patents, which covered all years
  - 9424 concepts in papers, 8471 concepts in patents
  - 19413 substances in papers, 27568 substances in patents
The concepts clustering presents the difference that the research fields of papers focus on the fundamental research, such as material structure and electrical properties, while the patents show the hot topics of application fields, such as batteries, semiconductors.
Different roles of substances in papers and patents

The roles of substances in papers are related to properties, while the patents focus more on the technical or engineered material use, industrial manufacture or additive use.
Link scientific research to industry

- **Data extraction and processing**
  - Title, abstract, concepts in papers or patents
  - Terms of industrial classification from International Standard Industrial Classification of all economic activities (ISIC)
  - Term Frequency - Inverse Document Frequency (TF-IDF) processing

- **Data mining**
  - Cosine similarity computing
  - Determine the threshold
  - Find the relationship between scientific research and industry, industry coverage, economic activities and industry focus.
The papers are linked to the industrial classification, such as the electrical equipment, the consumer electronics, on contrast, few patents are linked to these areas.
Conclusion

• We introduced literature research methods in knowledge discovery and proposed a knowledge discovery framework for XML-literature-data, which tailored for the CAplus database.

• We designed a customized tool for the CAplus data transformation, and XML data files were mapped into an internal processing file format.

• We presented the data mining methods to indicate the differences between the fundamental research and technology development, based on the same indexing system.
Future Work

• The customized tool that extracts the citation data.

• Add literature data mining methods:
  – Citation network analysis
  – Topic modeling for concepts
  – Deeper mining for substances and roles

• To study and understand the relationship between the fundamental research and technology development
References


• Le, S.S., Polytechnic, N. Technological innovation trend of graphene technology: A research based on the patentometric analysis. World nonferrous metals, 2017(9), 94–95.


• Zheng J. Comparative analysis of research paper and high level research paper of graphene field. Advanced materials industry, 2016(10), 48–51.
Thank you!

zoulx@mail.las.ac.cn