The search for R&D partners based on patent data

N Borodin¹, D Korobkin¹, A Bezruchenko¹ and S Fomenkov¹
¹Volgograd State Technical University, Volgograd, Russia

E-mail: dkorobkin80@mail.ru

Abstract. Nowadays, companies still rely heavily on expert knowledge when selecting research and development partners (R&D, R&D). In this paper, it is proposed to identify and select potential technology partners based on the analysis of the USPTO patent array using the Problem & Solution model. The process of identifying R&D partners will be based on the similarity of solutions to technological problems. This paper shows the development of software for extracting the semantic SAO structures from USPTO patent documents, which are used to form the Problem and Solution structures, and selecting potential technology partners based on them. The theoretical significance of the study lies in the developed algorithms for parsing patents; extracting SAO using semantic text analysis; formation of Problem-Solution structures based on SAO data; identifying R&D partners based on similarities in solutions to technological problems.

1. Introduction

A patent search is a procedure for selecting information on patents corresponding to a specific request, which can be carried out on one or several grounds. This is a time-consuming process that is often assigned to intellectual property professionals, but it is necessary - both for those who wish to obtain a patent and for those who wish to use an existing invention.

Patent searches [1-3] are currently carried out using special electronic information retrieval systems. The search is carried out on an array of patent documents or data. This is a rather narrow type of information search since it is carried out only based on patent documentation, which contains specific, structured data on a technical solution, possessing the necessary completeness and, most importantly, reliability, since it is verified as a result of an examination of the patentability of an invention.

There are several types of search, for example, personal (corporate) search, which is carried out by the name of the firms-owners of patents (corporate search by the applicant's name) or the names of inventors (author's search by the author's name).

Nowadays, finding R&D technology partners [4-7] requires the assistance of an expert in patent analysis. In turn, the examiner is required to use patent search, which is currently carried out using special electronic information retrieval systems. However, searching is a very laborious process due to the large number and complexity of electronic databases, each of which has its advantages and disadvantages. Conducting a search using electronic databases on the Internet is considered quite complete and objective, but it must be supplemented with a “manual” search through a detailed analysis of the information received.
In this paper, it is proposed to identify and select potential technology partners based on the analysis of the USPTO patent array using the Problem & Solution model. The process of identifying R&D partners will be based on the similarity of solutions to technological problems.

2. Analysis of search engines for patent documents

2.1 USPTO (uspto.gov)

This resource is a database of US patents, data which is stored in XML, pdf, tiff, etc. formats and is available for download.

The resource interface is quite simple, you need to find the category of files you are interested in and the years for which it is planned to obtain patents.

After downloading the patent data (in our case, the XML format) for the period of interest, you can proceed to the most difficult thing - finding the necessary information. The difficulty lies in the fact that the file contains a large number of patents (approximately 5000 in the file), which complicates the search process. The data in the file is strictly structured according to Standard St.36 Version 1.2 Recommendation for the processing of patent information using XML [8].

Thus, uspto.gov has a user-friendly interface, but the convenience of finding the required patent information is minimized due to the size and complexity of the XML document.

2.2 Google Patents

A product from Google that allows you to quickly and easily search for patent information, including the capabilities of named, numbering, and subject searches. Worldwide patents granted. It is possible to navigate and view full information about the patent of interest.

2.3 Yandex.Patents

The product of the Yandex almost completely repeats the functionality of an analog of Google, except for the possibility of searching for patents around the world. This resource only supports the search for patents of Russia and the Soviet Union.

2.4 Comparison of existing solutions

The results of the comparative analysis of the above solutions are presented in table 1.

| Criterion/solution   | uspto.gov | Google Patents | Yandex.Patents |
|---------------------|-----------|----------------|----------------|
| Actual data         | +         | +              | +              |
| Download patent     | +         | +              | +              |
| Named search        | -         | +              | +              |
| Number search       | -         | +              | +              |
| Subject search      | -         | +              | +              |
| View full patent information | + | +              |                |
| World patent database | -      | +              | -              |

3. An algorithm for extracting SAO structures

It is necessary to extract the "Subject-Action-Object" structures from the abstract and description texts. To solve this problem, an algorithm has been developed that consists of the following steps:

1) Splitting the text into sentences;
2) Splitting sentences into words;
3) Delete sentences containing stop words. This list was obtained by analyzing the texts of patent documents and contains words that occur in sentences that have no value when extracting SAO structures;
4) Removing the part of the offer that has no value when extracting the SAO;
5) Parsing a sentence with Stanza [9] (see Figure 1);
6) Getting from the data obtained at the last stage of the SAO structure, where:
   - VERB-Action;
   - nmod, obj, obl and their child elements-Object;
   - nsubj and its child elements are Subject.

Figure 1. Parsing a sentence with Stanza

4. **An algorithm for determining the similarity of technological problems**

The algorithm for finding problems similar to the user's request consists of the following steps:
   1) Parsing the sentence entered by the user using Stanza;
   2) Getting from the data obtained at the last stage of Action and Object;
   3) Search for entries in the SAO table where the action row matches the value highlighted in the previous step. When finding matches, it is assigned a coefficient of 1;
   4) Search for entries in the SAO table where the object string matches the value selected in step 2. When finding matches, it is assigned a coefficient of 1;
   5) Search by synonyms identified for action from the user's query using Word2Vec technology, the SAO table entries matching this field. When finding matches, it is assigned a coefficient of similarity of the context synonym to the source word;
   6) Search by synonyms identified for object from the user's query using Word2Vec technology, the SAO table entries matching this field. When finding matches, it is assigned a coefficient of similarity of the context synonym to the source word;
   7) Processing the list of found SAO records, when finding duplicate SAOs, their coefficients are summed up.

5. **Designing a technology partner search software**

Based on the analysis of existing solutions, it was decided to implement a client-oriented software that allows users to interact with the system to implement the goals of searching for research and development partners. In addition, the creation of a command-line interface for the administrator is proposed.

The projected (To-Be) process is designed to speed up the search for R&D partners and to save the resources of the stakeholder. The diagram of the automated process is shown in Figure 2.
Developing software will allow companies or individuals to get rid of an intermediary in the form of an expert when looking for technology partners in R&D, and experts, in turn, will speed up the process of finding potential like-minded people for clients.

The search will be performed based on the similarity of the technological problem indicated in the request and the problems previously prepared by the software from the patent array located in the SAO object database. The user will receive a list of similar problems and their solutions provided by other companies, based on which further interaction and partnership development are expected. Also, for the convenience of familiarization with patent information, it is supposed to implement the functionality for viewing the full-text description of the patent of interest on the problem.

Data for filling the database is added by parsing XML files from the USPTO database and extracting SAO structures from these files.

As a result of the analysis of the subject area, as well as by requirements set by the head, a list of requirements was formed for the developed software for searching for technological partners based on patent data:

- parsing of patent files;
- extraction of semantic structures "Subject-Action-Object" (SAO);
- formation of Problem-Solution structures;
- identification of R&D partners;
- visualization of presentation of technical problems, solutions, and R&D partners;
- viewing the full-text description of the patent.

Parsing of patent files is performed for files in XML format and is used to form a patent database. The extraction of the semantic structures "Subject-Action-Object" (SAO) comes from the Abstract and Description sections of the patent.

The Problem-Solution structures are generated based on the extracted SAO structures, where the problem is Action and Object, and the solution is Subject.

The identification of R&D partners is based on the similarity of solutions to technological problems.

Developing an interface to visualize the presence of technical problems, solutions, and R&D partners involves creating a page with a list of columns:

- the name of the company;
- patent number;
- title of the patent;
- problem;
- solution.
From the list it is meant the ability to go to the page with the full-text description of the patent, which includes the fields loaded from the database:
- the name of the company;
- patent number;
- title of the patent;
- authors;
- description;
- claims;
- abstract;

The implementation of the administrator interface implies interaction through the command line and allows you to parse patents, extract SAO, and form Problem-Solution structures.

User - a person who can search for technology R&D partners and view the full-text description of the patent. An administrator is a person who has access to a command-line interface with the ability to start parsing patents, retrieving SAO, and forming the "Problem-Solution" structure.

The above use case diagram lists the following functions:
- entering the name of a product or technology to find partners;
- viewing technology partners; transition to full-text description;
- parsing of patent files in XML format to form a patent database;
- extraction of semantic structures "Subject-Action-Object";
- formation of Problem-Solution structures based on extracted SAO structures;
- identifying R&D partners based on the similarity of solutions to technological problems;
- visualization of presentation of technical problems, solutions, and R&D partners.

6. Architecture of software

The architecture of the software [10] under development can be illustrated in the data flow diagram shown in Figure 3.

The software consists of the following blocks:
- The parser that receives the XML file of the patent as input;
- Module for extracting SAO structures from which "Problem-Solution" structures are formed;
- A module for finding technology partners that determine the similarity of problems;
- Web interface.

The server part of the application is implemented using the Django framework. All interaction with the server is reduced to 2 operations. This is a necessary and sufficient minimum, in a specific application implementation:
1) GET: getting data from the server (in our case, the Html code of the page is returned);
2) POST: sending data to the server.
For structuring and storing data, it was decided to use a PostgreSQL relational database. The database stores the entities of the patent information (Patents in Figure 4) extracted by the parser, and the SAO obtained at the stage of extracting the "Subject-Action-Object" structures.

The Patents table contains the following fields:
- "id" – a unique identifier of the patent in the database, which has the type guid;
- "author" - a string containing the authors of the patent;
- "title" - a string containing the name of the patent;
The SAO matrix contains the following fields:

- "id" – the unique identifier of the SAO;
- "subject" – subject;
- "action" – action;
- "object" – object;
- "patent_id" - a link to the patent from which the SAO is extracted;

The relationship between tables is one-to-many, since one SAO structure is extracted from one patent, but one patent can have multiple such structures.

An example of patent records is shown in Figure 5.

![Figure 5. Patent records](image)

An example of SAO records is shown in Figure 6.

![Figure 6. SAO records](image)

The client part of the application is implemented in Html using bootstrap and is responsible for displaying the user's request input line and a table of possible R & D partners, as well as for displaying detailed information about the patent.
7. The functionality of the software for technology partners searching

When the user enters the correct query and finds potential technology partners, the "Entered query" line will be filled in and a table with possible like-minded people will be drawn, consisting of columns:

- Company;
- Inventors;
- Name of the patent;
- Patent number;
- Problem.

The screen with the found partners is shown in Figure 7.

![Search for technology partners](image_url)

| Company                        | Inventors                                      | Name of the patent | Patent number | Problem                                                                 |
|--------------------------------|------------------------------------------------|--------------------|---------------|-------------------------------------------------------------------------|
| Sony Corporation               | Hiroshi Shimazu, Junichi Yokota, Ryogo          | Recording device,  | US07057760B1  | updates the free-capacity data stored in the nonvolatile memory so as to reduce the value of the free-capacity data stored in the nonvolatile memory by as much as the amount of the desired data recorded onto the recording medium |
| Toyota Jidosha                 | Yasushi Inoue, Katsuki Kaisha                   | Abnormality        | US07751946B1  | reduced Therefore, the calculation load or the memory capacity involved in the parameter identification are |
| Petrobras S.A. - Petrobras     | Vladimir Mate, Lisbina de Campos, Porti Cipriano, José De Medeiros, Jr. | Deep water high-capacity anchoring system and method of operation thereof | US07751946B1 | concerns an anchoring system by jetting applied to light anchors, with a high load capacity, |

![Figure 7. Potential R&D partners](image_url)

From the table, by clicking on the patent number, you can go to its full-text description, the page of which is shown in Figure 8, which contains the fields:

- Patent number;
- Name of the patent;
- Patent holder Company;
- Abstract;
- Description;
- Claims.
In addition, the application has a command-line interface for the administrator, who can perform the following actions:

- Parse the patents and write them to the database;
- Extract from the patent texts of the SAO structure;
- Retrain the Word2Vec model.

8. Acknowledgments

The reported study was funded by RFBR and Administration of the Volgograd region according to the research projects 19-47-340007, 19-41-340016.

References

[1] Wang Xuefeng, Ren Huichao, Chen Yun and etc 2019 Measuring patent similarity with SAO semantic analysis Scientometrics 121 10.1007/s11192-019-03191-z

[2] Korobkin D M, Fomenkov S A and Golovanchikov A B 2018 Method of identification of patent trends based on descriptions of technical functions Journal of Physics: Conference Series 1015 032065

[3] Park Hyunseo, Ree Jason and Kim Kwangsoo 2012 An SAO-based approach to patent evaluation using TRIZ evolution trends IEEE 6th International Conference on Management of Innovation and Technology, ICMIT 2012

[4] Kim Chulhyun and Lee Hakyeon 2020 A patent-based approach for the identification of technology-based service opportunities Computers & Industrial Engineering 144 106464

[5] Vasiliev S S, Korobkin D M, Kravets A G, Fomenkov S A and Kolesnikov S G 2020 Extraction of Cyber-Physical Systems Inventions’ Structural Elements of Russian-Language Patents Studies in Systems, Decision and Control 259 55-68

[6] Kharitonov A, Korobkin D M, Kravets A G, Fomenkov S A and Kolesnikov S G 2020 Extraction of Morphological Features of Technical Systems from Russian Patent CEUR Workshop Proceedings 2475 205-213

[7] Vayngolts I I, Korobkin D M, Fomenkov S A and Kolesnikov S G 2010 The Software and Information Complex Which Uses Structured Physical Knowledge for Technical Systems Design Communications in Computer and Information Science 1084 42-51

[8] Standard St. 36 Version 1.2 Recommendation for the processing of patent information using XML (Extensible Markup Language). - Introduction. 11/23/2007. - World: WIPO, 2007. - 14 pp
[9] Neural Pipeline  *Stanza: official site* https://stanfordnlp.github.io/stanza

[10] Python 3.7.10 documentation  *Python: official site* https://docs.python.org/3.7