CASE STUDY

Pharma Research Literature Analysis (PubMed)

Developed for pharmaceutical companies and research institutes by Megaputer Intelligence

Background

With the rapid development of biomedical and pharmaceutical research fields, the volume of published papers has increased dramatically. PubMed, the world's largest publicly available research database, alone has over 28 million literature citations. Thousands of records are being added to PubMed every day. It's overwhelming for researchers to keep up to date with both existing and new literature. They are seeking ways to automate the extraction of key information from literature.

A global pharmaceutical company was evaluating the possibility of stepping into the field of auto-immune diabetes and developing new drugs. It needed to analyze existing literature to find key information such as what clinical trials were conducted on auto-immune diabetes, what proteins and drugs are related to diabetes, which researchers are most productive and respected in this field.

Challenge

To gain a better understanding of the field, the company decided to study literature from the top biomedical and pharmaceutical database PubMed. In addition to extracting and normalizing entities, relations, and facts of interest, the company wanted to provide its analysts with the graphical and interactive presentations of the results to help them make better decisions.

A quick search for "diabetes" matched about 600,000 published citation records, including abstracts, online books, and links to full-text articles from PubMed Central database as well as from publishers' websites. It would be extremely difficult and time consuming for the company's research team to manually extract entities and facts of interest from all these data (full-text articles or even just abstracts). To make things worse, on average 50 to 100 more articles matching "diabetes" are added to PubMed daily. Text analysis tools the company tried using to automate the analysis, were capturing only key words and phrases and could not reliably extract more informative facts of practical interest. Neither manual analysis, nor traditional text analysis tools could cope with the task, and the company was seeking a more capable solution.

Solution

Megaputer used its proprietary data and text analysis software, PolyAnalyst™, to develop an automated literature analysis solution that extracts key information from text contents, discovers novel knowledge, and presents the findings to researchers and decision makers in easy to comprehend form.

The solution can read documents in any format. It supports several largest biomedical databases, such as PubMed and Cancer.gov. Any search queries against those databases can be configured in the solution itself. Using *diabetes* as a filter, all 600,000 articles were loaded into PolyAnalyst.

Megaputer
Your Knowledge Partner™

For more information:
Megaputer Intelligence | info@megaputer.com | 812-330-0110
www.megaputer.com

The solution's uses entity extraction capabilities for identifying related entities such as drugs and proteins by capturing linguistic and semantic relationships between these terms. For example, *CD28* is a protein that prevents auto-immune diabetes, and thus could be of great interest to researchers. To find out what can be activated by *CD28* in diabetes process, one creates a query extracting all sentences that contain *CD28* as the subject and synonyms of *activate* as the predicate. Then the list of all "objects" encountered in those sentences is compiled and normalized according to the industry standard ontologies such as MedDRA and MeSH. The resulting list contains everything activated by *CD28*: information of particular interest to the pharmaceutical company. In-depth linguistic analysis capabilities of the solution enable the researchers to accurately extract thousands of entities covering key information such as proteins, genes, chemicals, drugs, biological functions, and relations between them.

Comparing the shifts in the frequencies of extracted keywords by dates of publications and clinical trials, the company detected research trends: what topics were gaining popularity from one year to another. PolyAnalyst provides flexible graphical tools for presenting the discovered insights to researchers and decision makers in easy to comprehend form. This facilitates the broad use of extracted knowledge for practical decision making.

The solution provides social network analysis, in which the extracted researchers' names are cross-linked based on their collaboration in publications and clinical trials. The company can use this information to identify the most productive and well-connected researchers in this field and evaluate potential efficiency of partnering with academic institutions and other pharmaceutical companies.

The implementation of the Megaputer's text analysis solution enabled the team of researchers of the pharmaceutical company to grasp core knowledge related to auto-immune diabetes within days, reach out to several most productive research groups, and propose three pathways as the R&D phase directions.

Benefits

- **Easy literature access and query management.** Users can configure search queries and directly load data from resources such as PubMed. All analyzed literature contents are stored in the system.
- **High accuracy and throughput**. The solution provides high recall and precision of analysis of huge amounts of documents. It can be easily customized by modifying queries for different tasks.
- **Discovering unanticipated facts**. In addition to finding patterns researchers are specifically looking for, PolyAnalyst enables them to discover unanticipated facts of potential importance.
- **Increased efficiency.** The solution automates the process of information extraction and saves time and human efforts. Entities and facts of interest can be extracted from text contents in minutes.
- **Visualization and knowledge discovery**. Exploring key information with convenient graphical tools helps reveal previously hidden facts. This capability is especially useful for practitioners who have interest in new research domains.

Megaputer
Your Knowledge Partner™

For more information:
Megaputer Intelligence | info@megaputer.com | 812-330-0110
www.megaputer.com