UDK 004.654

Database for laboratories and industry

Federov N.S (ITMO University, Faculty of Software Engineering and Computer Systems)
Malyshev I.D. (ITMO University, Infochemistry Scientific Center),
Shcherbakova E.A. (ITMO University, Infochemistry Scientific Center)
Scientific supervisor - Skorb E.V. PhD (ITMO University, Infochemistry Scientific Center)

The computerization of chemical knowledge began relatively recently, near the end of the 20th century. Since then, the information has been gathered into massive databases containing chemical and physical parameters as well. Databases such as PubChem[1], Chemspider[2], ZINC[3], etc. contain only information about the structure of the substance and its characteristics. Databases like Reaxys[4], meanwhile, contain the reactions themselves as entities. Such databases are useful for scientists in planning syntheses. Regretfully, each of the databases has certain problems. Limited access, lack of proper filtering, repeats, no automatic filling are among them. However, the existing problems are solvable. This study proposes a model of an updated database, the number of disadvantages of which is minimized.

The database is represented as a bipartite graph: some nodes correspond to substances, others to reactions. It is not a trivial storage for data. The database supports basic graph algorithms. Deikstra algorithm, search in width, and search in depth are some of them. Such a representation advantage to keep all cause-and-effect relationships accounted for. This choice is also motivated by the fact that Bartosz Grzybowski's group of scientists got great results for their program Chematica[5] , which also has a bipartite graph structure . The most relevant formats for recording molecular structures were chosen. Among them are SMILES, IUPAC, InChi, MOL, XYZ etc. Information about already known compounds and reactions was downloaded using open sources. Unknown molecules expand the database due to the auto-complete module.

A database with a graph-based representation was projected. The proposed optimization of the database architecture will significantly increase the productivity of industrial and scientific laboratories. The future goal is to implement OPSIN [6] – program capable to convert one recording format to any other existing one in order to reduce the search time and save the memory resources. The disadvantages of developing a database with graph representation are high cost of maintaining and storing information.

Literature:

1. Qingliang Li, Tiejun Cheng, Yanli Wang, Stephen H. Bryant PubChem as a public resource for drug discovery // Drug Discovery Today 2010, 15: 23–24;
2. Harry E. Pence and Antony Williams ChemSpider: An Online Chemical Information Resource // J. Chem. Educ. 2010, 87(11):1123–1124;
3. John J. Irwin and Brian K. Shoichet ZINC − A Free Database of Commercially Available Compounds for Virtual Screening // J. Chem. Inf. Model. 2005, 45(1): 177–182;
4. Jonathan Goodman Computer Software Review: Reaxys // J. Chem. Inf. Model. 2009, 49(12):2897–2898;
5. Karol Molga, Sara Szymkuć, and Bartosz A. Grzybowski. Chemist Ex Machina: Advanced Synthesis Planning by Computers // Acc. Chem. Res. 2021, 54, 5, 1094-1106.

6. Lowe D. M. et al. Chemical name to structure: OPSIN, an open source solution // J. Chem. Inf. Model. 2011, 51(3): 739–753.