Другие разделы химической науки

Big data analysis in chemistry

Martynko E.A.[1], Afanaseva K.K.[2], Dmitriev V.A.[3]

[1] *Department of analytical chemistry*

[2] *Department of organic chemistry*

[3] *Laboratory of medicinal chemistry*

Every year the amount of published chemical documents increases. There are approximately 10 000 journals publishing articles in different fields of chemistry [1]. Every year, over 20 000 new compounds are described in these articles [2]. Together with journals, patents represent a valuable source of chemical information, but manual extraction of information from them is even more difficult due to large amount of tecnhical language and huge number of patents published every year [3].

Despite the diversity of various chemistry branches, end-users have common information demands: from finding papers of relevance for a particular chemical compound, family, or reaction (known as information retrieval, or IR) to extracting all chemicals or named chemical entities mentioned in a document (chemical entity recognition, CER), including associated information about chemical and physical properties, preparation, or toxicological data.

In information processing, a named entity is an abstract or physically existent real-world object, that can be denoted with a proper name [4]. In the field of chemical information retrieval systematic and trivial names and compound formulas are considered named entities.

Information retrieval (IR) is defined as finding the subset of those documents that satisfies a particular user demand within a large collection of documents [5]. User demands can be expressed as natural language text or as chemical structure. Chemical entity recognition (CER) refers to the process of automatic recognition of chemical entity mentions in text [6]. This approach is mainly used for database compilation and further analysis, e.g. QSAR (quantitative structure-activity relationship) modeling [7].

In this report, certain methods for data pre-processing and data extraction will be discussed. Successful information retrieval from articles and patents using text mining approach will be presented.

1. D. L. Roth. Chemical Information for Chemists: A Primer, (**2014**) 29-52;
2. Martin E. et al. Data Mining in Drug Discovery, (**2014**) 75-98;
3. Schneider N. et al. J. Med. Chem., 59 (**2016**) 4385-4402; IF 6.253
4. Zhang Y. et al. Database, 2016 (**2016**) 049; IF 3.755
5. Krallinger M. et al. Chem. Rev., 117 (**2017**) 7673-7761; IF 52.613
6. Akhondi S. A. et al. J. Cheminformatics, 7 (**2015**) S10; IF 3.893
7. Park S. et al. J. Chem. Inf. Model., 58 (**2018**) 244-251; IF 3.804