Chemical Structure Data Retrieval Algorithm for Chemistry Online Teaching

Guang-min Jiang[1(✉)], Xin Dai[2], and Wei Hou[2]

[1] Teng Zhou Vocational Education Centre, Teng Zhou 277500, China
[2] Zaozhuang Vocational College of Science and Technology, Zaozhuang 277500, China

Abstract. In the process of chemical network education, there are some problems in chemical molecular structure retrieval, such as low retrieval efficiency and slow retrieval speed, which can not meet the needs of teaching. Therefore, a large-scale chemical structure data retrieval algorithm is proposed for chemistry online teaching. Through the analysis of the chemical data, the chemical structure of the molecule was obtained. Using JSP technology and driver, the retrieval speed is improved. In the process of chemistry online teaching, large-scale chemical structure data can be retrieved. Through the comparative experiment, the retrieval speed and efficiency are taken as the experimental indexes. The retrieval speed ratio of this method is more than 2.3, and the retrieval time is about 100 s.

Keywords: Online teaching · Chemical structure · Data retrieva · Molecular structure

1 Introduction

Under the current novel coronavirus pneumonia, many schools are facing the reform of offline education mode. Online education has become the main trend of teaching nowadays. Among them, organic chemistry is an important branch of chemistry, which plays a very important role in the process of undergraduate talent training. Learning this course well is of great significance for the future study, work and scientific research of science and engineering students [1]. In the process of chemistry learning, the mastery of chemical molecular structure is an important link in the process of learning. In order to make online chemistry learning more vivid, it is necessary to search large-scale chemical structures for chemistry teaching. Therefore, it is urgent to develop a more efficient chemical structure data retrieval method.

With the sustainable development and progress of national economy, more and more researches are focused on the field of chemical structure data. JSP technology is a text-based, display centered development technology, which combines XML conversion format with database, and provides a separation mode of dynamic and static combination for the system [2]. In a typical database, the application of JSP program to the website can provide dynamic content for the system, and complete the tasks of database connection by using network template. Traditional methods for large-scale chemical

W. Fu et al. (Eds.): eLEOT 2021, LNICST 390, pp. 16–24, 2021.
https://doi.org/10.1007/978-3-030-84386-1_2

structure retrieval have some disadvantages [3], in reference [4], nine undergraduate chemistry students were investigated, and the challenges of online learning chemistry were described, including their omission of laboratory. Such as low efficiency and slow speed of data detection, which can not meet the needs of users.

In view of the above problems, a large-scale chemical structure data retrieval algorithm based on JSP technology is proposed. In order to shorten the waiting time of users, it is necessary to improve the speed of data retrieval. JSP technology is usually used to achieve large-scale data retrieval. Aiming at the problem of rapid growth of chemical structure data, the molecular structure processing mode is established. The experimental results show that the algorithm has fast retrieval speed and high efficiency, and can meet the needs of users.

2 Method

2.1 Data Analysis of Chemical Molecular Structure

The representation of molecular structure information is usually displayed in computer system by graphic representation, linear coding and structure coding. Due to the large space occupied by molecules in chemical structure, it is not suitable for large-scale chemical structure data retrieval [4, 5]. For the retrieval of chemical structure data, we need to use SDF format file to store the molecular form of chemical structure. Because SDF file storage format is connected by two-dimensional structure, it is usually suitable for molecular data storage and calculation in computer [6]. The information about molecular molar data in chemical structure mainly includes two parts: structure data information and physicochemical data information.

Due to the large amount of data contained in the molecular structure data information, it is more complex to retrieve the structure data. Therefore, it is necessary to search a large number of atoms and bonds in molecules. According to the historical records of physical and chemical properties in different databases, information such as ID, molecular formula, alias and water solubility can be retrieved. Taking the chemical molecular structure of dichloroacetic acid as an example, the molecular structure is shown in Fig. 1.

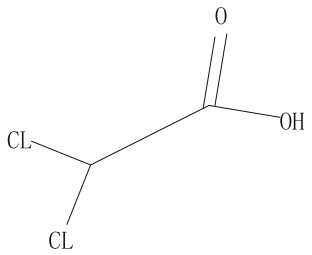


Fig.1. Chemical molecular structure of chloroacetic acid (two chloroacetic acid)

The structure of dichloroacetic acid is used as the basis for file storage, and the specific description information is shown in Table 1.

Table 1. Two chloroacetic acid format file

Molar data structure	Mrv0541	02241214202D				
	6	5	0	0	0	0.0000
	0.8225	1.4289	0.0000			
	0	0	0	0	0	
	1.2375	0.6542	0.0000			
	0	0	0	0	0	
	0.8113	0.0000	0.0000			
	0	0	0	0	0	
	1	2	1	0	0	0
	2	3	2	0	0	0
Physical and chemical data	><DRUGBANK_ID> DB08808 ><GENERIC_NAME> Dichloroacetic acid					

The end symbol of chemical structure data is "END", and the end symbol of physical and chemical data is "$$$$".

Before searching chemical molecular structure in chemistry online teaching, it is necessary to normalize the structure data [7–9]. In the process of structure matching, we need to use graph isomorphism algorithm, so sometimes we will encounter the problem of inconsistent retrieval structure: for example, when the user enters the question structure shown in Fig. 2, this problem will occur. From a chemical point of view, formula (1) and formula (2) in Fig. 2 express the same molecule. But from a graph point of view, they are two different graphs. In this way, if users submit queries with different structure diagrams, they will get different results, which should be avoided in chemical structure queries. Therefore, the data of chemical structure were normalized as follows.

(1) (2)

Fig. 2. Different drawing methods of the same molecular structure

Before loading data to database, use structure conversion function to structures.sdf The structure data in the file is converted one by one. The method is as follows: first, read

the 2D structure data of a compound from the structures [10–13]. SDF file, and construct the corresponding molecule object with the smile string. Then the member function of the Molecule. object is called aromatize (true) to detect the aromaticity of the Molecule structure and handle it accordingly. Finally, the other member function toformat () of the molecule object is used to convert the molecule into a mol format string and a smiles string, respectively. The normalized structure data of the compounds were encoded to form the values of each field in the data table molecue, which were imported into the database.

2.2 Molecular Data Retrieval Based on JSP Technology

The molecular structure of a compound is a two-dimensional undirected connected graph. The molecular structure retrieval is realized by transforming the molecular structure retrieval into the graph isomorphism problem [14, 15]. Substructure matching has been proved to be NP complete problem, that is, the time consumption of the algorithm increases exponentially with the number of nodes (atoms). In the process of searching chemical teaching resources, when JSP technology is used for database operation, it usually has the following operation steps: connection, query, output and result display [16–18]. The interface of connection class is usually used to connect database in JSP, and it is implemented by various drivers. Generally, the Prepared Statement is used to query the database and save the query results to the system. The main feature of this method is that there is no preprocessing stage. The sliding window is always moved back one bit. The comparison order of characters in the pattern is not limited. It can be from front to back or from back to front.

The driver provided by Microsoft is used to hold the query and update of data, and the type information of Prepared Statementcan be viewed from the code. According to the programming model, for large-scale data parallel computing, the main task is decomposition and result merging [9, 10]. First, the information in the function is assigned to the key and value to form the corresponding key value list as the input stage. In each input phase, any frame belongs to the value combined by the same key. Then the value is assigned to each node, waiting for the next task. Because the running function needs to be processed, therefore, in the last node, the generated results should be merged to complete the result set [19, 20].

Based on JSP technology, the efficient retrieval ability mainly depends on the allocation stage. Suppose that the total number of parallel issues in the processing stage of Map task and Reduce task is M and R respectively. Suppose there are s calculation nodes A1, A2, A3…, As-1, As. The specific calculation formulas (1) and (2) are shown below.

$$M = \sum_{i=1}^{s} M_i \tag{1}$$

$$\mathrm{R} = \sum_{i=1}^{s} R_i \tag{2}$$

From formula (1) and formula (2), it can be seen that M_i and R_i are values on different nodes, and distributed processing can promote the matching of chemical structure data with high speed.

3 Simulation Experiment and Analysis

3.1 Data Source

In this experiment, the molfile of MDL company is used as the question map, and the SDF file is used as the collection of the target map. It is widely used, and its experimental data can be obtained free of charge on the NCI open chemical database website.

3.2 Experimental Environment

The experimental distributed environment needs four hosts, whose frequency size is 2.00 GHz and memory size is 3 GB. The simulation software is matlab (2019 a). The configuration information of each node in the computer is shown in Table 2.

Table 2. Node information configuration

Number	IP	Role	Software environment
1	218.195.2555.44	NameNode: JobTracker	Ubuntu11.05, JRE1.5, Hadoop0.10
2	218.195.2555.45	DataNode: TaskTracker	
3	218.195.2555.46		
4	218.195.2555.47		

In order to verify the rationality of the large-scale chemical structure data retrieval algorithm based on JSP technology, the following experiments are carried out using retrieval speed and retrieval efficiency as experimental indicators.

3.3 Experimental Results and Analysis

Retrieval Speed Verification

The target set of the experiment is to compare the retrieval rate of the traditional algorithm (reference [3] method) and the proposed algorithm in the system for large-scale chemical structure data retrieval under the condition of consistent data, combined with the retrieval algorithm. In the experiment, the selected data size is 100 M, with the size of 10 atoms as a molecular data structure, increasing from 0 M to 6 M of N atoms. The experimental comparison results are shown in Fig. 3.

When a heteroatom is added to large-scale chemical structure data, the time of data retrieval can be reduced. With the increase of the number of heteroatoms in chemical structure, the time of data retrieval for chemical structure tends to be stable. As can be seen from Fig. 3, the time spent on data retrieval using traditional algorithms fluctuates back and forth in 200–250 s with the increase of the number of atoms. However, with the increase of the number of atoms, the retrieval time of this algorithm is gradually reduced, and finally stabilized at about 100 s.

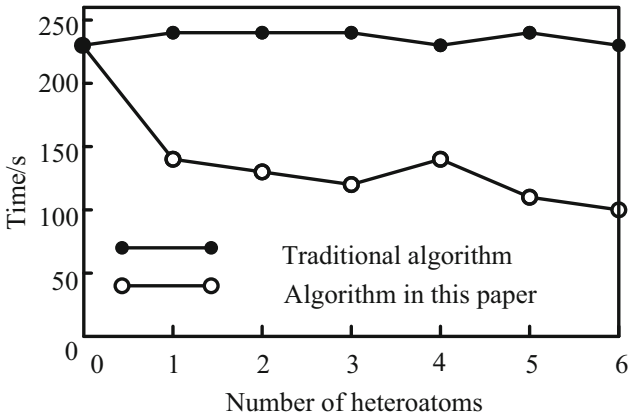


Fig. 3. Search rate comparison results of two algorithms

Retrieval Efficiency Verification

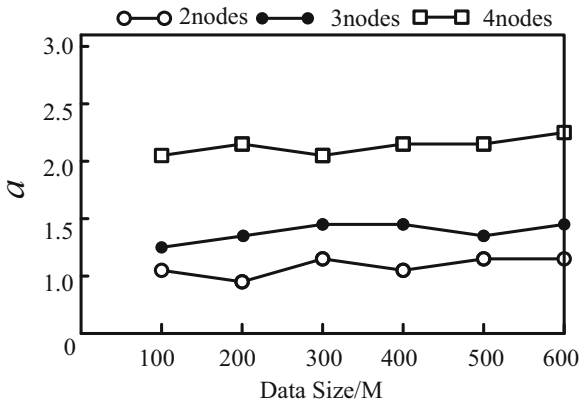
To ensure the same amount of data, the traditional algorithm and the algorithm in this paper are used to retrieve the large-scale chemical structure data in the distributed mode. The experimental data set is A = {100, 200, 300, 400, 500, 600}, typical chemical structures were selected for experimental verification. With the increase of the amount of data, the retrieval speed of distributed compound matching is greatly improved. The search speed based on distributed can reflect the advantage of great difference, and the definition of speedup is shown in formula (1).

$$a = \frac{T_S}{T_P} \tag{3}$$

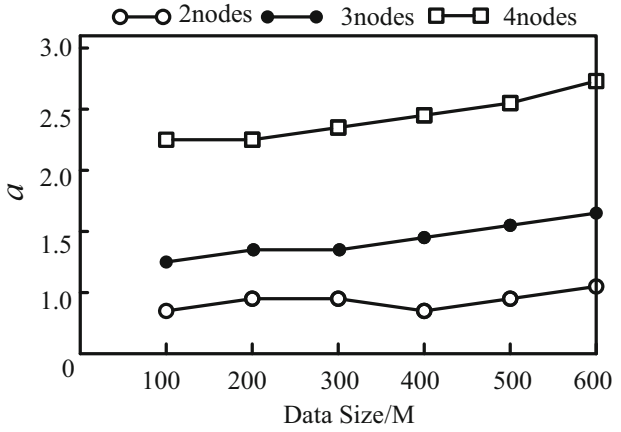
According to the Formula (3), T_S is the time spent by the program executing the retrieval on the system, T_P is the time spent by the program executing the retrieval in parallel on the distributed cluster, and a is the speedup ratio. As the amount of data continues to grow, the time of chemical structure data matching increases linearly. The traditional algorithm and the algorithm in this paper are used for data retrieval of large-scale chemical structures in this trend, and the acceleration is shown in Fig. 4.

It can be seen from Fig. 4 that with the increase of the number of nodes, the speedup ratio of chemical structure data retrieval increases in a certain proportion. When the number of nodes is 2, the speedup ratio of traditional retrieval algorithm is 1.15, and that of this algorithm is 1.25. When the number of nodes is 3, the speedup ratio of traditional retrieval algorithm is 1.49, and that of this algorithm is 1.65. When the number of nodes is 4, the speedup ratio of traditional retrieval algorithm is 2.35, and that of this algorithm is 2.85.

After the above experimental contents, the following experimental conclusions can be drawn. With the increase of the number of atoms, the time of data retrieval using traditional algorithms fluctuates back and forth in 200–250 s. However, with the increase of the number of atoms, the retrieval time of this algorithm is gradually reduced, and finally stabilized at about 100s. When the number of nodes is 2, the speedup ratio



(a) Traditional algorithm



(b) Algorithm in this paper

Fig. 4. Acceleration ratio

of traditional retrieval algorithm is 1.15, and that of this algorithm is 1.25. When the number of nodes is 3, the speedup ratio of traditional retrieval algorithm is 1.49, and that of this algorithm is 1.65. When the number of nodes is 4, the speedup ratio of traditional retrieval algorithm is 2.35, and that of this algorithm is 2.85. It can be seen that the large-scale chemical structure data retrieval algorithm based on JSP technology has stable scalability, can meet the requirements of large-scale data retrieval, and the retrieval speed is fast.

4 Conclusion

In the case of chemistry online teaching, aiming at the problem of low efficiency of chemical structure data retrieval, a large-scale chemical structure data retrieval algorithm based on JSP technology is proposed. Using JSP technology, large-scale chemical structure data retrieval can be realized through the network. Experimental results show that the algorithm has fast retrieval speed, high efficiency and strong scalability. It can be better applied in chemistry online teaching, which is conducive to improving the teaching effect.

References

1. Liu, D., Wang, J.: Design of Internet online auxiliary teaching system based on Web. Mod. Electron. Tech. **40**(20), 28–30 (2017)
2. Guo, H., Li, Y., An, H.: A Parallel communication algorithm in supersonic COIL's calculations using multiblock mesh. J. Comput. Res. Dev. **53**(5), 1166–1172 (2016)
3. Lee, M.W.: Online teaching of chemistry during the period of COVID-19: experience at a national university in Korea. J. Chem. Educ. **97**(9), 2834–2838 (2020)
4. Jeffery, K.A., Bauer, C.F.: Students' responses to emergency remote online teaching reveal critical factors for all teaching. J. Chem. Educ. **97**(9), 2472–2485 (2020)
5. Xing, C., Xiong, Z., Li, Y., et al.: Construction algorithm of geometric invariant based on area ratio. Appl. Res. Comput. **34**(6), 1900–1904 (2017)
6. Lin, Z., Shuai, J.: Multi-crossover strategy of multi-objective cellular genetic algorithm research on flexible job-shop scheduling problem. Sci. Technol. Eng. **17**(7), 69–76 (2017)
7. Liu, S., Glowatz, M., Zappatore, M., et al. (eds.): e-Learning, e-Education, and Online Training, pp. 1–374. Springer, Heidelberg (2018)
8. Harshman, J., Yezierski, E.: Assessment data-driven inquiry: a review of how to use assessment results to inform chemistry teaching. Sci. Educ. **25**(2), 97–107 (2017)
9. Penny, M.R., Cao, Z.J., Patel, B., et al.: Three-dimensional printing of a scalable molecular model and orbital kit for organic chemistry teaching and learning. J. Chem. Educ. **94**(9), 1265–1271 (2017)
10. Peng, Z., Jimenez, J.L.: KinSim: a research-grade, user-friendly, visual kinetics simulator for chemical-kinetics and environmental-chemistry teaching. **96**(4), 806–811 (2019)
11. Liu, S., Lu, M., Li, H., et al.: Prediction of gene expression patterns with generalized linear regression model. Front. Genet. **10**, 120 (2019)
12. Liu, S., Li, Z., Zhang, Y., et al.: Introduction of key problems in long-distance learning and training. Mob. Netw. Appl. **24**(1), 1–4 (2019)
13. Guo, Y., Gao, N.: Rational synthetic parameter analysis of open-framework AlPOs Based on Data Mining Method. Chin. J. Inorgan. Chem. **32**(3), 457–463 (2016)
14. Cooper, M.M., Stowe, R.L.: Chemistry education research—from personal empiricism to evidence, theory, and informed practice. Chem. Rev. **118**(12), 6053–6087 (2018)
15. Shapiro, H.B., Lee, C.H., Roth, N.E.W., et al.: Understanding the massive open online course (MOOC) student experience: an examination of attitudes, motivations, and barriers. Comput. Educ. **110**, 35–50 (2017)
16. Zancanaro, A., Nunes, C.S., de Souza, D.M.J.C.: Evaluation of free platforms for delivery of Massive Open Online Courses (MOOCs). Turk. Online J. Dist. Educ. **18**(1), 166–181 (2017)
17. Pölloth, B., Teikmane, I., Schwarzer, S., et al.: Development of a modular online video library for the introductory organic chemistry laboratory. J. Chem. Educ. **97**(2), 338–343 (2019)

18. Crimmins, M.T., Midkiff, B.: High structure active learning pedagogy for the teaching of organic chemistry: assessing the impact on academic outcomes. J. Chem. Educ. **94**(4), 429–438 (2017)
19. Krijtenburg-Lewerissa, K., Pol, H.J., Brinkman, A., et al.: Insights into teaching quantum mechanics in secondary and lower undergraduate education. Phys. Rev. Phys. Educ. Res. **13**(1), 010109 (2017)
20. Gadaleta, D., Lombardo, A., Toma, C., Benfenati, E.: A new semi-automated workflow for chemical data retrieval and quality checking for modeling applications. J. Cheminf. **10**(1), 1–13 (2018). https://doi.org/10.1186/s13321-018-0315-6