AN EXPLORATION OF CDIO-BASED BIG DATA ANALYTICS AND APPLICATIONS COURSE FOR CHEMISTRY STUDENTS

Dazhou Li[1] and Wei Gao[2*]

[1]Dr., Shenyang University of Chemical Technology, China, lidazhou@syuct.edu.cn
[2]Prof. Dr., Shenyang University of Chemical Technology, China, gaowei_syuct@foxmail.com
*Corresponding Author

Abstract

Ability to solve realistic problems in chemical big data, the CDIO model inherits and develops the concept of the significant reform of engineering education in Europe and the United States for more than 20 years, and nearly 40 universities in China have become the demonstration university of China's CDIO engineering education model. China CDIO Engineering Education Model has officially accredited Shenyang University of Chemical Technology's Computer Science Technology College in 2020. This article is a review of Shenyang University of Chemical Technology's Computer Science Technology College in the process of China's CDIO Engineering Education Model Certification. Universal experience with CDIO engineering education has shown that CDIO standards not only contribute to enhancing the quality of education but also provide an opportunity to improve the quality of engineering education. A good foundation is laid for the systematic development of education. To this end, we introduced the CDIO engineering education model into the Shenyang University of Chemical Technology's big data analytics and applications, and in the design of pragmatic course teaching. We apply the CDIO concept to the human workforce training program, course system construction for chemistry majors, and faculty of Computer Science Technology majors. A series of investigations and practices have been carried out. To address the problem of offering big data analytics and applications courses for chemistry students, we introduce the CDIO engineering education model into the content, teaching methods, and teacher preparation are all aspects of this project. We have given a "problem-oriented" for both theoretical teaching and practical teaching, considering the learning situation of chemistry students. Based on learning with an inspiring interactive theoretical teaching method, "result-oriented" practical teaching mode, and project design-oriented, students learn professional knowledge and engineering thinking through practical engineering projects and improve their engineering awareness and practical skills. Finally, through the combination of theory and practice, chemistry majors can understand the theories and methods of big data analysis and applications, as well as the principles and methods of big data analysis and processing.

Keywords: CDIO, engineering education, big data analytics and applications, chemistry majors, talent training mode

1. INTRODUCTION

The application of big data creates the value of big data (Wu, 2013, p.97). With the rapid development of information technology, the application of the results of big data technology has been integrated into all walks of life (Chen, 2014, p.171). Big data technology in chemistry is embodied in the production,

distribution, research, and management of modern chemistry (Sagiroglu, 2013, p.42). Convenience, but also can promote the efficient allocation of resources in the production process (Chen, 2014, p.275).

As a new resource element of modern chemistry, chemical big data not only leads to the construction of chemical modernization but also drives the chemical supply side (Lusher, 2014, p.859). With the rapid development of the information age, the essential socio-economic value of chemical big data is more prominent (Hu, 2017, p.179). Chemical big data is used to build smart chemistry. In the innovation of chemical production and operation methods, chemical entrepreneurship and innovation, and scientific decision-making in the chemical industry, chemical big data has been used for the construction of intelligent chemistry through its good (Gibb, 2013, p.248). The chemical big data has explosive growth with powerful vitality. It has become a powerful engine for the advancement of modern chemistry, through promoting the development of modern chemistry (Yeguas, 2014, p.389).

The Shenyang University of Chemical Technology, a chemistry university that trains chemistry professionals, aims to train "complex" qualified professionals. How to make use of its advantages? How to innovate the model and approach of training chemistry professionals and how to teach traditional teaching in light of the actual situation of the construction of big data in chemistry? What is the way of reform in concepts and methods? The above are essential issues in the reform and development of the major at the Shenyang University of Chemical Technology.

Big Data Analytics and Applications is a core foundation course for Big Data Science Technology majors. As a chemistry school, we have offered Big Data Analysis and Applications Course for the students from the Computer Science Technology major and Big Data Science major. However, the Big Data course offering for chemistry students is still in the exploratory stage. Compared with other science and engineering colleges, chemistry colleges have significant chemical characteristics in terms of specialization and other aspects. Therefore, the teaching of big data analysis and application courses should be closely associated with the training plans and goals of chemistry students.

On the other hand, chemistry colleges and universities have only a few years of experience in establishing Big Data Science technology majors, and big data analysis and application courses are still in the exploratory stage. The curriculum is still being continuously revised and improved. Therefore, before offering a course on Big Data Analysis and Applications to chemistry students, it is necessary to review the teaching methods, curriculum, teaching content, experimental design, and course assessment. In outcome-oriented teaching design, it explains how to identify learning outcomes, develop training programs and prepare syllabuses, emphasizing reconstructing training programs - optimizing knowledge structures and clarifying three relationships; in outcome-oriented teaching implementation, it explains how to reform training models, reform classroom teaching, and develop innovative education, emphasizing on CDIO's one vision, one syllabus, one standard and one set of strategies. In the evaluation of result-oriented education, the course explains what a quality assurance system is, how to build it, and how to carry out the continuous improvement.

Notably, as a course for undergraduate students in chemistry, the content of the Big Data Analytics and Applications course is related to the chemistry big data, in which the integration is not deep enough, and the teaching characteristics of chemistry colleges have not been reflected. Therefore, this paper combines the actual situation of the Shenyang University of Chemical Technology in the development of big data in chemistry with the actual teaching effort of big data analysis for chemistry students. The study is an exploration of teaching methods with applied courses. The research content has reference value for the construction of the Big Data Science technology curriculum system with characteristics of chemistry colleges.

2. DESIGN OF TEACHING CONTENT OF BIG DATA COURSE FOR CHEMISTRY MAJOR

2.1. Use of Globally Accessible Large Chemical Databases

At this stage, chemistry research is inseparable from the support of a sizeable public chemistry database, which is already a pain point of chemistry education at this stage. Students can not use an extensive free chemistry database, which means that they are not able to enter the forefront of chemistry research in the world. Public databases, such as PubChem (Kim, 2016, p.1202), BindingDB (Liu, 2007, p.198), and ChEMBL (Bento, 2014, p.1083), represent large public domain compound activity databases. As a teaching content, this aspect can be considered and drawn on the Internet crawler technology in the current big data application practice course (Barbosa, 2007, p.441). The access interface, provided by the Python language and the R language through the public database to quickly access, search, and download the interesting chemical content in the database, also needs to be taught. Only by fully mining the chemical data in the

database can the frontier chemical information be mastered and used.

2.2. Data Visualization and Evaluation of Chemical Space

The first step in chemical data analysis is usually the visualization and compact representation of millions of compounds, which is also a significant challenge for big data analysis. The solution is usually accomplished by projecting an extensive collection of compounds into a low-dimensional space, which is convenient for visual inspection and intuitive analysis of the human brain. It can help detect chemical structures with new chemical scaffolds and physicochemical properties, evaluate different libraries, and determine chemical spatial regions with specific pharmacological characteristics. Typical methods include principal component analysis, constructed topographic maps, Kohoning networks, diffusion maps, and interactive maps obtained by projecting high-dimensional descriptor spaces.

The number of potential molecular structures that can be enumerated in theory is enormous. For example, the database GDB-17 contains 166.4 billion molecules, which are based on simple rules of chemical properties and synthetic feasibility (Ruddigkeit, 2012, p.2864). There can be up to 17 possible combinations of C, N, O, S, and halogen atoms. Even a fast algorithm that can process 100,000 molecules per minute requires three years of calculations to annotate the full GDB-17. These data sets even pose new challenges to the analysis of traditional chemical compound collections. As a teaching content, this aspect can be considered and drawn on the distributed computing technology in the current big data application practice. Big data parallel computing frameworks, such as Hadoop and Spark, should be taught (Zhao, 2011, p.343, Chen, 2016, p.919). The advantages of the cloud computing platform can break the bottleneck of limited local stand-alone computing capabilities in chemical computing. Only in this way can the parallelization and cloudization of calculations be implemented, thereby speeding up the calculation process of chemical analysis. Pymol is an easy-to-use, powerful software for visualizing molecules as well as proteins, developed by Schrödinger. Researchers can request the latest educational version from the official website, while the open-source version of Pymol, which can be downloaded directly from the website, is older. So, choose the version to download according to the needs.

2.3. New Pharmaceutical Design Methods

The new drug design requires searching for a sizeable virtual compound database. When we search for a vast virtual chemical space, it is necessary to combine efficient search and multi-parameter optimization strategies to filter out molecules with sub-optimal properties as early as possible. For example, physicochemical and synthetic feasibility filters can be placed in front to reduce the number of compounds. Another strategy is reaction-driven fragment-based redesign. Based on known chemical reactions and industry available components, through the usual multi-step and multi-parameter optimization process, the search for candidate compounds that meet specific properties to generate chemically diverse and synthetically feasible compounds can be implemented.

These reaction-based methods have been successfully applied to design new biologically active compounds. Therefore, these methods still maintain the chemical space of the model while developing new chemical structures. As a teaching content, this aspect can be considered on the massive search and ultra-large-scale distributed database technology in the current big data computing practice course. The usage of big data storage databases, such as Hive and HBase should be taught (Poggi, 2017, p.55, Wang, 2014, p.71). The cloud computing platform can avoid the problem of low search performance of the existing local stand-alone data storage platform. Drawing on the basic concepts of Google, Baidu, and other big data search engines, a fast search method for sizeable virtual compound databases can be implemented.

3. CURRICULUM TEACHING METHODS

3.1. Subjects and Objectives of the Training

For chemistry students to become a qualified person with big data analytics skills, they need to have the following foundations: (1) The complete fundamentals of mathematics, such as advanced mathematics, probability theory, mathematical statistics, and linear algebra. The above courses provide students with theoretical knowledge of the subject and an understanding of the mathematical models used in processing data. (2) Data collection skills include traditional sampling surveys and acquire data from the Internet. (3) Necessary computational programming skills, such as the ability to master computational tools such as Hadoop and Hive in the large-scale multi-source heterogeneous data processing.

In particular, the most compatible scripting languages, Python and R, should be mastered by chemistry students. In the course of big data analysis and applications, chemistry majors can use Python and R to perform scientific computation. Visual mapping and data pre-processing, as well as practical problems in

chemical big data using algorithms such as clustering, regression, and classification, are also the tools that should be mastered with Python and R to analyze and model. Ultimately, the chemistry students combine theoretical knowledge with practical application to lay the foundation for a future career in chemical big data analysis and mining research by Python and R.

3.2. Use of Heuristic Interactive Theory Teaching

In the process of teaching theoretical content in the curriculum, a strictly lecture-based approach is likely to lead to a dull classroom atmosphere, no interest, and low involvement. Interest is the best teacher. First, students' interest in learning is stimulated. Then, inspiring interactive teaching is adopted. Students actively participate in the teaching process, so that the teaching concept of "teacher-led to student-led" can be realized. In other words, theoretical teaching in the classroom starts with questions. These questions create learning situations and stimulate students' interest. Students actively participate in the process of problem analysis and thinking. In this process, students consciously want to learn and understand the theories and algorithms that are used to solve tasks and problems. In this way, the excellent teaching effect of setting doubts and stimulating learning with interest is achieved.

Generally speaking, there are three main advantages of using heuristic teaching methods in theoretical teaching: (1) students' interest in learning and desire for knowledge are stimulated; (2) students think actively according to the problems and tasks, and their engagement in the learning process is mobilized; (3) the traditional teaching method of filling in the gaps between students and teachers is changed, and how teachers ask students' questions is changed to how students ask questions and think individually.

Besides, for a specific practical problem, other ways of teaching, such as student lectures or group discussions, can also be adopted. Students are assigned to prepare for the content of the class in advance. Multiple students take turns teaching the content that has been prepared in advance in the classroom. When we were employing the interactive theoretical teaching method, the current situation can be effectively changed. The "hard" and "reluctant" state of the students can be changed to a "happy" and "eager" state. The "unidirectional indoctrination" is changed to "multi-directional contact" between teachers and students as well as between students. Classroom teaching is energetic and fun. Theoretical teaching content is no longer boring but exciting.

3.3. Establishment of a New Teaching Model for Student Self-Learning

With compressed teaching, students have much free time. If students' spare time is fully utilized and their interest in the curriculum is enhanced, they become more active rather than passive learners. That's what we're mainly trying to do now. How do students ask questions? The ability of students to access appropriate materials in conjunction with the appropriate equipment is also the objective of the primary pedagogical reform. The proportion of regular student grades in the performance assessment system can be increased. The understanding of painful chemistry points can be improved by assigning discussion questions, asking relevant questions in class, and exploring severe problems after class for students. The mainline of teaching content is to "lighten theory, promote skill development, and enhance the application. "

How do we teach students to use databases? Students cannot only develop ideas about learning problems but also to ask questions. Models are developed and elevated to theory, which is the primary method of analyzing problems in the field of chemistry. Rather than formulas for fundamental theorems, chemistry students are expected to learn in college so that they can fully understand how their predecessors have learned from a large number of Identify relevant patterns in the test data.

3.4. Adoption of the "Result-Oriented" Practical Teaching Model

In the process of practical teaching, the design of experimental project content should be tightly integrated with the theoretical teaching content. Therefore, the "result-oriented" teaching method can be used in the practical education process. The core of the concept of "result-oriented" education is that through the design of teaching approaches and the realization of objectives, the students can achieve significant results after learning through practical education. We bring this concept into the experimental program of the Big Data Analytics and Applications course. The experimental content and objectives are designed based on the practical problems and tasks that are introduced in the theoretical teaching process in the relevant algorithms. Then it is used as the basis for solving practical problems and tasks. Finally, the students' experiments are analyzed to find the gap between them and the objectives of practical teaching. The introduction of "result-oriented" practical teaching methods is conducive to the change from knowledge-oriented to result-oriented. Correspondingly, the course assessment changes from the traditional single paper-based test to multiple assessment methods. In the practical teaching of big data analysis and application courses, students fully grasp the process of analyzing data, analyzing data, and analyzing data in

the process of completing experimental projects—a method for solving problems to complete relevant primary research. OBE emphasizes expanded opportunities, i.e. learning outcomes as a guide, assessment results as a basis for modifying, adapting, and flexibly responding to student learning requirements when appropriate. Outcomes prevail rather than certificates. Traditional education students gain certificates are oriented to the completion of a prescribed course of study for a specified period in terms of credits, and instructional design begins with the construction of a curriculum to determine the appropriateness of meeting the course's instructional objectives. Inverse design starts with the needs, which determine the training objectives, which in turn determine the graduation requirements, which in turn determine the curriculum.

4. UPGRADING TEACHERS

Big data analysis and processing in chemistry is a practical application of big data science theory, technology, and methods in chemistry, and related chemical-related fields. We use the unique facilities of chemistry to enhance contact and cooperation between teachers and other faculty members. Through the application of big data analytics techniques in chemistry, we realize the construction and development of big data in chemistry. Big data analytics and application courses are typically taught by computer-related faculty. However, the course content for chemistry students is mainly about the application and processing of big data in chemistry, so the professional chemistry knowledge must be enhanced. At the Shenyang University of Chemical Technology, cooperation and exchange between interdisciplinary teachers are strengthened. For example, during winter and summer vacations, computer teachers associated with big data analysis and application courses participate in courses on foundations of chemistry, applications of chemistry, and medications—professional faculty training for core courses such as manufacturing methods. The diverse educational cultures and learning experiences of computer science teachers are elevated for teaching courses in big data analytics and application.

5. CONCLUSIONS

To address the issue of providing big data analysis and application courses for chemistry majors in chemistry colleges, we focus on the content, teaching methods, and faculty. A preliminary exploration of the preparatory aspects was conducted. In the case of chemistry majors, the theoretical and practical aspects of the course are given as "problem-based learning" with inspiring interactive theoretical teaching methods and "result-oriented" practical teaching mode. Finally, through the combination of theory and practice, chemistry majors can understand the principles and methods of big data analysis and processing, as well as the principles and methods of big data analysis and processing. The ability to master the computer techniques to analyze and solve practical problems in chemical big data can be obtained.

6. ACKNOWLEDGEMENT

We would like to take this opportunity to acknowledge the hard work of all our editors and also the efforts of all the anonymous reviewers who have provided very valuable and helpful comments. We thank them all for their invaluable contributions, and for helping with our research. Project supported by The Science and Technology Funds from Liaoning Education Department (No. LQ2017008); Doctoral Research Startup Fund Project of Liaoning Province (No. 2016011968); Liaoning Higher Education Society 13th Five-Year Plan General Topics (GHYB160163); Ministry of Education, Department of Higher Education, Collaborative Education Project with University-Industry Cooperation (201801128005, 201902233001); Shenyang University of Chemical Technology Education and Training Project (No. 35).

REFERENCE LIST

Barbosa, L., & Freire, J. (2007, May). An adaptive crawler for locating hidden-web entry points. In Proceedings of the 16th international conference on World Wide Web (pp. 441-450).

Bento, A. P., Gaulton, A., Hersey, A., Bellis, L. J., Chambers, J., Davies, M., ... & Nowotka, M. (2014). The ChEMBL bioactivity database: an update. Nucleic acids research, 42(D1), D1083-D1090.

Chen, M., Mao, S., & Liu, Y. (2014). Big data: A survey. Mobile networks and applications, 19(2), 171-209.

Chen, C. P., & Zhang, C. Y. (2014). Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. Information sciences, 275, 314-347.

Chen, J., Li, K., Tang, Z., Bilal, K., Yu, S., Weng, C., & Li, K. (2016). A parallel random forest algorithm for big data in a spark cloud computing environment. IEEE Transactions on Parallel and Distributed

Systems, 28(4), 919-933.

Gibb, B. C. (2013). Big (chemistry) data. Nature chemistry, 5(4), 248-249.

Hu, Y., & Bajorath, J. (2017). Entering the 'big data'era in medicinal chemistry: molecular promiscuity analysis revisited. Future science OA, 3(2), FSO179.

Kim, S., Thiessen, P. A., Bolton, E. E., Chen, J., Fu, G., Gindulyte, A., ... & Wang, J. (2016). PubChem substance and compound databases. Nucleic acids research, 44(D1), D1202-D1213.

Liu, T., Lin, Y., Wen, X., Jorissen, R. N., & Gilson, M. K. (2007). BindingDB: a web-accessible database of experimentally determined protein–ligand binding affinities. Nucleic acids research, 35(suppl_1), D198-D201.

Lusher, S. J., McGuire, R., van Schaik, R. C., Nicholson, C. D., & de Vlieg, J. (2014). Data-driven medicinal chemistry in the era of big data. Drug discovery today, 19(7), 859-868.

Poggi, N., Montero, A., & Carrera, D. (2017, August). Characterizing bigbench queries, hive, and spark in multi-cloud environments. In Technology Conference on Performance Evaluation and Benchmarking (pp. 55-74). Springer, Cham.

Ruddigkeit, L., Van Deursen, R., Blum, L. C., & Reymond, J. L. (2012). Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17. Journal of chemical information and modeling, 52(11), 2864-2875.

Sagiroglu, S., & Sinanc, D. (2013, May). Big data: A review. In 2013 international conference on collaboration technologies and systems (CTS) (pp. 42-47). IEEE.

Wang, H., Li, J., Zhang, H., & Zhou, Y. (2014, March). Benchmarking replication and consistency strategies in cloud serving databases: Hbase and cassandra. In Workshop on Big Data Benchmarks, Performance Optimization, and Emerging Hardware (pp. 71-82). Springer, Cham.

Wu, X., Zhu, X., Wu, G. Q., & Ding, W. (2013). Data mining with big data. IEEE transactions on knowledge and data engineering, 26(1), 97-107.

Yeguas, V., & Casado, R. (2014, August). Big data issues in computational chemistry. In Proceedings of the 2014 International Conference on Future Internet of Things and Cloud (pp. 389-392).

Zhao, H., Wang, Y., & Yang, L. (2011, October). Research on distance education based on cloud computing. In 2011 6th International Conference on Pervasive Computing and Applications (pp. 343-348). IEEE.