A software service outsourcing talent cultivation mechanism, based on big data processing technology

Jiping Li, Yaoming Ding & Zhimin Li

Hubei Engineering University
Xiaogan, China

ABSTRACT: A talent cultivation mechanism (TCM) plays an important role in cultivating high-quality talent in a university. In this article, the drawbacks first are highlighted of the current TCM for software service outsourcing (SSO) in local universities of China and, then, discussed is the necessity of using big data technology to improve the software service outsourcing talent cultivation mechanism (SSO-TCM). To this end, a data-driven SSO-TCM scheme is proposed, based on Hadoop, a programming framework that is open source and Java-based. It is suggested that off-campus information be acquired, such as the comments and suggestions from SSO enterprises, alumni and parents of students. Also the on-campus information should be acquired, such as students' comments and suggestions about the course. Through big data technology, the proposed scheme can be used to improve the SSO-TCM. Guidelines are provided for the design and development of a data-driven SSO-TCM, based on the Hadoop platform.

INTRODUCTION

In recent years, the global software service outsourcing (SSO) industry has developed rapidly. As reported in *Business Week*, one third of the world's software production is through outsourcing with an average annual growth rate of 29.2%. As a result, SSO enterprise demand for software development talent is rising sharply. In China, software industry revenues were said to have exceeded RMB 4 trillion by 2015, with an annual growth rate of more than 25% [1].

In 2013, China's software industry reported revenues of RMB 3.1 trillion, increasing by 24.6% year on year, which was much higher than the global average of 5.7% [2]. However, the objectives of education at local universities in China do not take account of market demand. This means graduates from local universities do not meet SSO industry requirements in terms of knowledge, personal ability and quality. Hence, there is an apparent contradiction between job hunting being difficult and a serious lack of SSO talent.

Many scholars have carried out research on ways to improve SSO talent cultivation [3-8]. However, all of them tend toward recommending an equilibrium between theoretical knowledge and practical applications, with a balance between general knowledge and specific expertise [9]. To balance theory and practice, two main methods are suggested.

The first method is CDIO (conceive, design, implement, operate). The second method involves co-operation between a university and an enterprise. The CDIO-based engineering education concept aims to cultivate IT professional talent [9-11], while co-operation between universities and enterprises aims to train students with applied practical skills [12].

For SSO the focus is on a talent cultivation mechanism (TCM), course syllabus, and experiments based on experience. But, this is not properly related to theory or engineering project practice, nor does it keep up-to-date with technology and the demand for SSO talent. Therefore, current SSO-TCM teaching lacks pertinent objectives that are kept up-to-date.

However, the advent of the *big data* era provides new opportunities to solve existing problems with SSO-TCM. Increasingly popular social media applications, such as Twitter, Facebook, blogs and on-line product reviews, lie within the big data spectrum and contain relevant information for business decision making [13]. Big data concepts and analytics can be applied to a variety of higher education administration and instructional applications, including recruitment, admissions processing, financial planning, donor tracking and student performance monitoring [14].

Local universities, as the main component of higher education of China, should make full use of the best-available and most cost-effective big data processing technology to help reform SSO-TCM, with the aim of cultivating SSO talent to meet market demand.

PROBLEMS OF SSO-TCMs AT LOCAL UNVERSITIES IN CHINA

Educational Objectives

Currently, there are two extremes in the objectives for talent cultivation at local universities in China, i.e., being *too high* or *two low*. The former pays more attention to theory and, similar to top-ranking universities, tends to put too much emphasis on system design, while neglecting practical ability. By comparison, the latter emphasises practical ability, while neglecting theory. The *too low* extreme results in graduates who *know how to operate, but not how to design*.

Obviously, this TCM approach is employment-oriented. But, the local university is very different from research-oriented top-ranking universities and also from employment-oriented vocational and technical colleges. The educational objectives of local universities are to cultivate application talent. That is to say, the practical ability of graduates from local universities should be much better than that of graduates from top-ranking universities, and their knowledge and ability should surpass those from vocational and technical colleges.

Drawbacks of *Basic Courses before Professional* in the TCM

Professional foundation courses mainly are scheduled in the first and second years at China's local universities, with professional courses beginning in the third. Most of the professional courses are held in the fourth year, and it is the busiest of the four years of higher education. However, in the first term of the fourth year, students are busy at job-hunting, examinations for certificates, preparing for the postgraduate entrance examination and working in enterprises.

As a result, the course attendance rate in the fourth year is very low. Moreover, due to working in enterprises, the scheduled study time for theory and experiments is very limited. All this tends to separate the study of theory from practical experience. So, even though a student may have graduated from a local university and become a member of an SSO enterprise, the student's ability may not meet the technical requirements of the SSO enterprise and, hence, the student may require special technical training.

SSO-TCM Technology Lags behind SSO Enterprise Technology

Although some local universities have carried out *university-enterprise co-operative projects*, their depth and breadth usually are insufficient due to a lack of investigation of the marketplace. Regular and effective communication between local universities and SSO enterprises is scarce, making it difficult to guarantee the required rapid and frequent upgrading of the university software development systems. The phenomenon, whereby, a new software-development technology is soon replaced, makes it difficult for the SSO-TCM of a local university to adapt to the demands of the software marketplace. In addition, teaching materials usually lag behind SSO technology, because of the rapid development of the technology. Hence, local university SSO-TCMs do not well-equip students for the workplace after graduation.

SSO Enterprise Talent Types

The SSO enterprise has a pyramid-type structure. Usually, senior architects are at the top of the pyramid; senior engineers in charge of projects, technology and product management lower down; and programmers are at the bottom of the pyramid [15]. The SSO graduates normally lack practical experience in software development. So, merely based on the theoretical knowledge and experiments in school, it is difficult for graduates to move into the upper levels of the SSO enterprise pyramid or become a top talent of the SSO enterprise. That is to say, the graduate can be only a bottom of the pyramid programmer, and so cannot meet the SSO enterprise demand for high-level talent.

BIG DATA AND BIG DATA SOURCES USED IN DECISION ANALYSIS

Characteristics of Big Data

The term big data usually refers to the accumulated data and information related to some industry. The data refers not only to the data of one type, but to data of various types, such as structured, semi-structured and unstructured [16]. According to a survey in *Computer World*, unstructured data accounts for 70% to 80% of all data in an organisation and is growing ten to 50 times faster than structured data [17].

Big data is not simply a pointless accumulation of a large amount of data; it includes the discovery of the direct or indirect relationships between those data. Big data allows the discovery and in-depth understanding of new, hidden values. Big data has three aspects: a) the data are numerous; b) the data may not be stored in a regular relational database; and c) the data are generated, captured and processed very quickly [18].

SSO-TCM Big Data Sources

The sources of SSO-TCM big data used for decision-making are various, but mainly include the following:

- Direct face-to-face feedback, such as students' comments and suggestions in theory and practice teaching; comments and suggestions from alumni and their parents; other off-campus opinions and suggestions;
- Collected information from Web pages, such as skills in demand by SSO enterprises; the content of SSO enterprise internal skills training, teaching material and training methods information; previous graduates' feedback about course-selection presented in the BBS (bulletin board system); curriculum evaluation and suggestions in the post bar;
- Feedback about theory and practice teaching, curriculum reviews and learning experiences from mobile social networks, such as MSN, Skype, Facebook, Twitter, blog, microblog, WeChat and QQ space;
- Students' library-access information, for example, download information, electronic documents, reference information;
- Interactive information on excellent course platforms, such as massive open on-line courses (MOOCs) and university open on-line courses (UOOCs).

Decision-Making based on Big Data Technology

The aim is to make effective decisions to adjust the SSO-TCM, so as to improve local university SSO graduates' software development ability. The feedback information from the SSO enterprise, student feedback (Facebook, Skype, Twitter, MSN, Renren, QQ space); interactive information between the teacher and students in the MOOC or UOOC platforms, etc, are all useful for the decision-maker.

Using the data, techniques, such as associate rules, machine learning, data mining, cluster analysis, text analysis and time series analysis can be used to find information to help the decision maker to identify weaknesses in the current SSO-TCM. Moreover, the data analysis not only potentially identifies the weaknesses in the current SSO-TCM, but also reveals the misunderstandings or difficulties students experience in the learning. With the extracted information, teachers may adjust the teaching content and methods, which will benefit graduates' knowledge and programming ability.

The Ecosystem of Apache Hadoop

Hadoop is a distributed, powerful, efficient, highly reliable, cost-effective computing system, which has the ability to store and handle petabyte-level mass data. It is currently used for cloud computing by some computing pioneers, such as Yahoo, Amazon, Facebook and eBay.

Hadoop is mainly composed of the HDFS (Hadoop distributed file system), MapReduce, Hbase, Pig, Hive, Zookeeper, Sqoop and Mahout. The most common components are HDFS and MapReduce. The HDFS is used to provide a highly reliable, high performance, scalable database system with real-time read/write operations. MapReduce supports parallel computing for massive datasets, with data partitioning and computing task scheduling. Data and computing are co-located for system optimisation, and there is good error detection and recovery.

Mahout is a library for machine-learning and data mining. Pig, a data stream processing platform, is used to generate a high-level scripting language and to operate a run-time platform that enables users to execute MapReduce. Pig is also used to load data, transform data formats and to store the final results, supporting optimal MapReduce operation. Hive, acting as a data warehouse, is used to add data to HDFS and to query data using a language similar to SQL (structured query language). Flume is a distributed, highly reliable, high-availability massive log aggregation system, used to collect and move large amounts of log data [19].

Sqoop is a data exporting tool between an RDBMS (relational database management system) and HDFS, and is used for relational data extraction, transformation and loading. Zookeeper is used to provide configuration maintenance, name services, distributed synchronisation and group services.

DATA-DRIVEN SSO-TCM, BASED ON HADOOP

Big data technology provides the ability to capture and store large volumes of public and private data together with the decision-making support to analyse the data [20]. The technology can be used to mine, identify, organise and extract implied information in a collection of structured, semi-structured and unstructured data.

A framework for a data-driven SSO-TCM for a local university was developed based on the open-source, cost-effective Hadoop platform. By using the unified data warehouse, Hive, as data storage, Hadoop can be used to analyse and process data at the bottom layer. By using the data mining module, Hadoop can call various services through the transparent interface to the top application layer facilitating real-time processing of the data.

The framework of a data-driven SSO-TCM based on Hadoop includes data acquisition, big data processing, data mining, results analysis and storage, as well as SSO talent cultivation scheme adjustment modules. The Hadoop-based framework for data-driven SSO-TCM is shown in Figure 1.

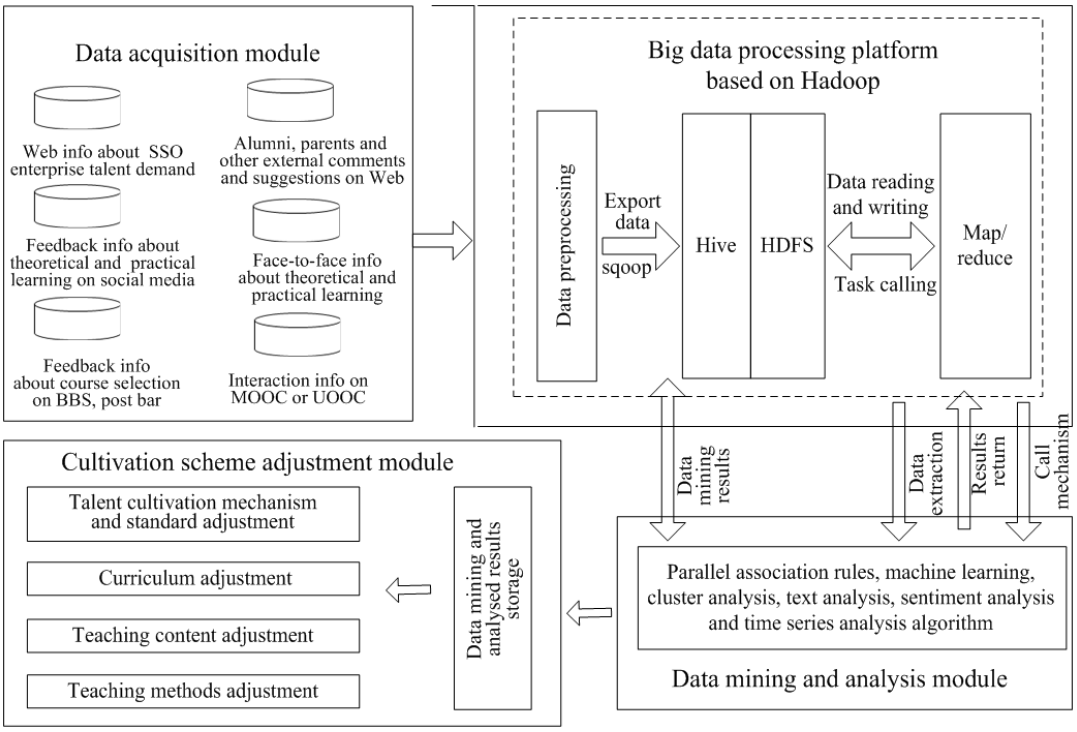

Figure 1: The framework of data-driven SSO-TCM based on Hadoop.

Data acquisition module

The data acquisition module is used to gather, filter and clean data so as to ensure that only the high-value fragments of data are passed to the data warehouse. The acquisition methods and support technology are shown in Table 1.

Table 1: Acquisition methods and support technology.

Data category		Acquisition method	Support technology
Market demand for SSO talent	Number of required SSO talent	Web crawler tools (Spider, Robot)	Batch crawling, distributed crawling, incremental crawling, scoped crawling, deep Web crawling
	Skills in specific programming language		
	Programming ability		
	Foreign language ability		
Direct feedback information	Face-to-face feedback information re theory study	Snaptrends - social media data collection tool	Keyword filters, word cloud, mood and sentiment analysis
	Face-to-face feedback information re on and off campus practical study		
	Interactive information between teachers and students via social networks	Video, audio, image and GPS sensors embedded in mobile device	
	Interactive information between teachers and students via MOOC or UOOC platform		
Indirect feedback information	Comments or suggestions left on social networks, email and post bar	Web crawler, video, audio, image and GPS sensors embedded in mobile device	Batch crawling, distributed crawling, incremental crawling, scoped crawling, deep Web crawling
	Comments or suggestions from email, BBS or post bar		
	Content of SSO-enterprise training information		
	Job-hunting information related to required professional ability		
	On-line public access catalogue (OPAC) logs, borrowing or database accessing records	Log file and records on library servers	Common log file format (NCSA), Extended log format (W3C), IIS log format (Microsoft)

The acquired data includes:

- SSO talent demand, SSO enterprise internal skills training content, teaching material, training methods;
- Alumni and parents' feedback information, external feedback information about courses;

- Readers' OPAC (online public access catalogue) log information, borrowing information (e.g. name of book, frequency of library use, stay time), database browsed and download information (e.g. name of the database, title of article, browse or download) and reference information;
- Feedback information about students' study in Facebook, Twitter, blog, microblog, WeChat, QQ space, BBS and post bar;
- Students' interactive information on MOOC and UOOC.

Big Data Processing Module

The big data processing platform is composed of a data pre-processing module, data warehouse Hive, HDFS and Map-Reduce. The data pre-processing module is used to transform acquired data into transaction data for data mining using association rules. The raw data are stored in Hive using the data migration tool, Sqoop. The HDFS and MapReduce - core components of the data processing module - are responsible for data storage and computing services, respectively. The HDFS and MapReduce run on the same node, which allows MapReduce to distribute tasks onto nodes with the required stored data, so as to efficiently parallel-process across the whole networked cluster. Hence, the computational overhead does not significantly grow with an increase in the volume of acquired data [21].

Data mining and analysis module

The data mining and analysis module is one of the core modules of a data driven SSO-TCM. By using association rules, machine learning, data mining, cluster analysis, text analysis and time series analysis, the data mining and analysis module can extract useful information from the cleaned data [20][22]. Details of the data mining and analysis module are shown in Table 2 below.

Table 2: Data mining and analysis module.

Data mining and analysis category	Supported technology	Application
Relationship analysis	Association rule learning	Recommendation system
Complex pattern analysis	Machine learning	Intelligent decision system
Combination statistics and machine learning with database management	Data mining	Market demand analysis
Unsupervised machine learning	Cluster analysis	Micro analysis
Large text collection analysis	Text analysis	Extract information from emails, Web pages
Social media analysis	Sentiment analysis	Extract information from social media
Analysing sequences of data points	Time series analysis	Forecasting technology or market demand trend

Adjustment Module for SSO-TCM

By using association rules, useful information can be extracted from the data. This extracted information can be used to adjust the curriculum, teaching content and teaching methods. This adjustment of SSO-TCM is the result of an objective data analysis, rather than being subjective; it also reflects the forecast demand for future professional skills and talent by the SSO industry.

DATA-DRIVEN SSO-TCM PROCESS

As referred to earlier, the information Hadoop uses includes SSO enterprise talent demand; internal staff training by SSO enterprises; students' feedback about theory and practice teaching; comments and suggestions about the curriculum; the proportion of teaching hours for theory and practice; comments from BBS and post bar; students' opinions and suggestions about teaching (e.g. Facebook, Twitter, blog, microblog, WeChat, QQ space); students' OPAC log (books borrowed, electronic resources browsed and downloaded); interactive information from MOOC and UOOC platforms. After data pre-processing, the data are stored in Hbase.

Data cleaning, extraction, integration, organisation and analysis of relationships between data are undertaken to extract the key information of skills required by the SSO enterprise and the skills an SSO enterprise should have. Hence, the components of a professional course to provide those skills can be determined. Accordingly, the existing curriculum, teaching content, and teaching method can be adjusted, so as to make the graduates from local universities better suited to the market, thereby, improving their employability.

CONCLUSIONS

A talent cultivation mechanism plays an important role in producing high-quality talent at universities. In this article, an adjustable data-driven software service outsourcing talent cultivation mechanism (SSO-TCM) is proposed as an intelligent decision-making system, based on Hadoop. By acquiring on- and off-campus information and using big data technology, the proposed scheme could transform education decision-making and practice. According to the extracted information from Hadoop, deep reforms of SSO talent cultivation at the university could be carried out. These would involve the curriculum, teaching content and teaching methods, so as to produce high-quality SSO talent with good professional skills.

ACKNOWLEDGEMENTS

The authors extend grateful thanks for the helpful suggestions of reviewers. This work was founded by the 2015 Key Research Topics of Hubei Provincial Educational Science Planning under Grant No. 2015GA038, and partly supported by both the National Science Foundation of China under Grant No. 61370223 and the Natural Science Foundation of Hubei Province under Grant No. 2014CFB577.

REFERENCES

1. Research in China (2013), China Computer Software Industry Report 2013, 20 September 2016, www.prnewswire.com/news-releases/china-software-industry-report-2013-235099641.html.
2. Research in China (2014), China Computer Software Industry Report 2014-2017, 20 September 2016, www.rnrmarketresearch.com/china-computer-software-industry-report-2014-2017-market-report.html.
3. Yang, L., Zhang, W., Chen, Z. and Zhang, Y., Practice of undergraduate's specialty construction for software service outsourcing direction. *China Electric Power Educ.*, 11, 55-56 (2014).
4. Dai, J. and Peng, J., Research and practice on personnel training mode of software service outsourcing. *Computer Knowledge and Technol.*, 10, **1**, 105-106 (2014).
5. Dai, J. and Yuan, H., Software service outsourcing course system construction based on school-enterprise cooperation. *Computer Knowledge and Technol.*, 10, **19**, 4486-4487 (2014).
6. Wang, X., Research on software service outsourcing talent cultivating model based on CDIO. *Educ. Teaching Forum*, 2, 204-205 (2015).
7. Ding, N. and Gu, Y., Research on the problems existing in teaching management of software service outsourcing talent cultivating mode. *China Electric Power Educ.*, 12, 27-28 (2014).
8. Zhang, H., Liu, S., Zhang, S. and Guo, H., Research on the problems and counter-measures of software outsourcing talents cultivating in colleges and universities. *Computer Educ.*, 9, 72-75 (2014).
9. Zhang, F., Zhang, Y., Ai, X. and Li, X., Research on a *2+1+1* IT professional talent training mode based on the CDIO engineering education concept. *World Trans. on Engng. and Technol. Educ.*, 12, **2**, 186-190 (2014).
10. Cao, W., Liu, Z. and Zheng, L., Reforming the teaching of a single chip microprocessor course based on CDIO engineering education. *World Trans. on Engng. and Technol. Educ.*, 11, **4**, 428-433 (2013).
11. Jia, S. and Yang, C., Teaching software testing based on CDIO. *World Trans. on Engng. and Technol. Educ*, 11, **4** , 476-479 (2013).
12. Gu, Q., Wang, X., Wu, Z. and Hua, L., Exploration and practice of college-enterprise co-operation talent cultivating in computer science at local universities. *World Trans. on Engng. and Technol. Educ.*, 12, **1**, 20-25 (2014).
13. Goh, T.T. and Sun, P-C., Teaching social media analytics: an assessment based on natural disaster postings. *J. of Infor. System Educ.*, 26, **1**, 27-36 (2015).
14. Picciano, A.G., the Evolution of Big Data and learning analytics in American higher education. *J. of Asynchronous Learning Networks*, 16, **3**, 9-20 (2012).
15. Liu, C.J., Zou, H. and Zou, N., Research on international applied software service outsourcing talents training mode. *Computer Educ.*, 22, 18-22 (2011).
16. Che, D., Safran, M. and Peng, Z., *From Big Data to Big Data Mining: Challenges, Issues, and Opportunities, in Database Systems for Advanced Applications*. Berlin: Springer, 1-15 (2013).
17. Gartner, Top 10 Strategic Technologies for 2010, 20 September 2016, www.gartner.com/newsroom/id/1210613.
18. Khan, N., Yaqoob, I., Hashem, I.A.T., Inayat, Z., Ali, W.K.M., Alam, M., Shiraz, M. and Gani, A., Big Data: survey, technologies, opportunities, and challenges. *The Scientific World J.*, 712826-712826 (2014).
19. Cavanillas, J.M., Curry, E. and Wahlster, W., *New Horizons for a Data-Driven Economy a Roadmap for Usage and Exploitation of Big Data in Europe*. Berlin: Springer, 48-49 (2015).
20. Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C. and Byers, A.H., Big Data: the Next Frontier for Innovation, Competition, and Productivity. Tech. Rep (2011).
21. Xie, G. and Luo, S., Study on application of MapReduce model based on Hadoop. *Microcomputer & Its Applications*, 8, 4-7 (2010).
22. Emani, C.K., Cullot, N. and Nicolle, C., Understandable Big Data: a survey. *Computer Science Review*, 17, 70-81 (2015).