International Journal of Engineering Education Vol. 35, No. 4, pp. 1110–1115, 2019
Printed in Great Britain

0949-149X/91 $3.00+0.00

Extraction of Core Competencies for Big Data: Implications for Competency-Based Engineering Education*

FATIH GURCAN
Karadeniz Technical University, Department of Computer Engineering, 61080, Trabzon, Turkey. E-mail: fgurcan@ktu.edu.tr

Big data industry is an innovative and dynamic working environment based on highly qualified workforce. As the big data phenomenon advances, the demands of the industry for the workforce having these skills and competencies have increased considerably in recent times. Accordingly, the engineering education programs today need to adapt these skills and competencies into their programs. Focusing on this issue, this study aims to extract the core competencies in-demand by the industry. These competencies are the critical ones to better guide the curriculum developers of the engineering education programs. The methodology of the study is based on topic modeling analysis of online job advertisements using Latent Dirichlet Allocation, a generative approach for probabilistic topic models, to automatically discover the trending topics in big data jobs. As a result, domain-specific competencies, developer competencies, soft competencies, business-oriented competencies and analytical competencies are discovered, which revealed that big data competencies contain a wide spectrum of knowledge domains and skill sets based on a multidisciplinary background. The findings of the study are very critical to guide the industry, academia, and big data communities for bridging the gap between the requirements of the industry and the engineering education programs.

Keywords: big data competencies; competency-based engineering education; big data curriculum; topic modeling; Latent Dirichlet Allocation

1. Introduction

Big Data is one of the key drivers profoundly changing the IT industry and playing crucial role in many digital transformations. Although there is no fixed definition of big data, in its broadest sense, big data can be defined as the different types of high volume data (texts, images, videos, etc.) generated by different sources (social networks, mobile devices, internet of things, etc.) processed and modeled with up-to-date analysis methods [1–3]. The introduction of big data technologies has led to a great enhancement in business operations and decision making processes. From this perspective, big data is one of the most valuable assets with strategic priorities for today's industries [1, 3, 4]. Big data-oriented analysis and applications provide significant contributions and insights for industries. In today's dynamic industrial environments, based on global competition, resources and investments allocated to big data research and practice are increasing day by day [1–5].

The big data field is a dynamic working environment in which actors, roles, responsibilities and competencies are frequently changing [6–8]. Working in this active field may be considered a dream job by most, because it offers incredible opportunities in terms of career, earning and dynamic work settings [4, 5, 9]. As the number of big data-based products and services increases, the demand for qualified big data professionals is ever-increasing over time. On the other hand, the number of qualified big data professionals does not increase in the same way as industry needs. Research indicates that skills gap and lack of qualified staff are the leading barriers to big data phenomenon [1, 4, 6]. It is estimated that the businesses will have significant challenges in the near future in finding and employing the big data professionals capable of meeting their needs [1, 6]. In this setting, the academia is inadequate to provide the necessary supply and contribution in terms of meeting the emerging needs and demands of the big data industry. This existing gap between industry needs and academic preparation was highlighted in many researches [1, 6, 10, 11].

In this context, the education of big data professionals with necessary qualifications to meet industry needs is an open question for academia [1, 5, 6, 10, 11]. In terms of an ever-developing big data phenomenon, a competency-based engineering education (CBEE) consistent with industry needs should be addressed in a collaborative manner by the industry and academia [12–16]. The methodology of CBEE is based on individual competencies (knowledge, skills, and abilities) different from the time-based education. The CBEE approach aims to gain students the capability to use and apply knowledge and skills acquired at the end of the education. Each specific skill or ability, known as a competency, is represented by a single component instead of a course or module. During the education process, a flexible outline is provided to students that enables them to progress independently of time, place, and pace of learning. Due to its specified

* Accepted 7 April 2019.

characteristics, CBEE offers more effective solutions than traditional time-based education especially for IT-oriented engineering disciplines so as to bridge the skill gap between industry and academia [12, 17]. From this point of view, revealing the competency requirements for big data may provide a better understanding for industry and academia [18–21].

Within this framework, the aim of this study is to identify the core competencies needed in the big data industry and to obtain implications that will set goals for competency-based education. In accordance with this aim, a topic modeling-based content analysis was performed on big data job ads using Latent Dirichlet Allocation (LDA) [22], a generative approach for probabilistic topic models [23]. The novelty of this study is that it was totally performed by LDA-based topic modeling in a fine granularity level. As a result of this experimental study, five competency domains reflecting essential knowledge and skills for big data were revealed. Our research is expected to contribute to industry, academia, and big data communities in bridging the gap between the requirements of the industry and the engineering education programs.

2. Data and methods

2.1 Data collection and preprocessing

Online job ads are an up-to-date information source that provides an important insight into the industry needs and trends. Therefore, this empirical analysis was performed on the big data job ads collected from indeed.com, a global employment platform offering advanced search-query options for employers and employees [24]. With the purpose of improving the consistency of the data, only the jobs ads including "big data" phrase in the title were extracted. When creating the data set, no filtering was applied for any country or employer, and all of the big data-driven job ads were added to the dataset. At the end of this process, the empirical dataset consisting of the 2175 job ads was created. The dataset covered a time period of six months, from February 2018 to June 2018.

In next phase, the preprocessing tasks were implemented on the unstructured dataset. In text mining, preprocessing is commonly used to structure and clean the textual data [25, 26]. The preprocessing tasks implemented in this context consisted of the following sequential steps. Initially, the tokenization was applied to separate the texts into simple tokens (words). The misleading words, special characters, punctuation and links were then removed. Afterward, stop words that have a high frequency in English texts (is, and, a, the, of, for, etc.) were deleted with the aim of helping in the generating meaningful topics. In the preprocessing phase, no stemming process was performed in order to avoid any loss of sense because the textual data consist entirely of technical words [25, 26]. After the preprocessing tasks, the document-term matrix (DTM) was created so as to employ the topic modeling on the dataset. A DTM is a numerical matrix that reveals the frequency of the terms that occur in a collection of documents. Each row denotes a document in the collection and each column denotes a unique word in the document in the DTM. The DTM built for the analysis consisted of 2175 rows and 9354 columns, meaning that 2175 text documents (job ads) were described by a word list consisting of 9354 terms.

2.2 Probabilistic topic modeling

Topic modeling is a probabilistic model used to identify the latent semantic structures, called as topics, in a large corpus. Latent Dirichlet Allocation (LDA) is a generative model commonly used in probabilistic topic modeling. The "latent" term in this model indicates the hidden semantic structures (topics) in the documents. The generative model refers the representation of the words in the documents as latent variables in an iterative probabilistic process based on Dirichlet distribution and thus the semantic modeling of the documents in this way [22, 23, 27].

For the implementation of the LDA model, the study used the MALLET topic model package [28]. The Mallet package is based on Gibbs sampling algorithm, used for hyper parameter optimization in training of the LDA model [29]. The package was employed with 1200 Gibbs sampling iterations. And also, hyper parameters of Dirichlet distribution were used with values of $\beta = 0.1$ and $\alpha = 50/T$ in the experiments [27]. The number of topics, represented by T, is a prediction parameter used to identify the granularity level of the topics. The Gibbs sampling algorithm with varied T-values (between 30 and 100) was employed so as to evaluate the results of the changing number of topics (T). At last, in a finer granularity level, the desired results were achieved when the number of topics was equal to 60. Following the LDA implementation, the discovered topics were interpreted using topics' keywords and ratings, and the labels were assigned to the topics. [23, 27].

3. Results

As a result of this analysis performed by LDA-based topic modeling, the 60 trending topics were discovered in a finer granularity level. In the following step of the analysis, the topics were associated with the basic disciplines and a competency taxonomy was

Table 1. Topics related to domain-specific competencies

Topic Label	Descriptive Keywords	Rate %
Educational background	science computer related engineering degree field equivalent preferred master	2.57
Experience	experience year hands relevant minimum recent demonstrable prior progressive	2.56
Big data platforms	hadoop hive spark hbase pig mapreduce hdfs sqoop kafka flume	2.33
Database technologies	database sql nosql relational server oracle cassandra mongodb including storage	2.12
Streaming data processing	spark hadoop kafka processing frameworks streaming apache storm building	2.04
Roles and duties	architect job developer position senior description roles qualifications duties	1.94
Real-time processing	time full real processes stack implement build batch ingestion billions	1.80
Open-source platforms	technologies hadoop source open platforms cloudera hortonworks latest github	1.80
Background requirements	required include background clearance candidates ability training face basics	1.79
Big data infrastructures	data big engineer principal developer infrastructures workflows building set	1.76
Azure services	data azure warehouse lake structured microsoft unstructured analytics suite	1.64
System management	system security management cluster admin operating configuration service	1.62
Data modeling	data modeling fitting segment design complex database graph display	1.61
Deep domain knowledge	experience knowledge understanding strong technical expert subject solid field	1.60
Data mining	data processing pipelines big architectures algorithms mining metadata extract	1.59
Data scaling and integration	scaling data integration big technology requirements design module system	1.51
Big data sources	data big sources stations understand industry clients internal assets	1.45
Big data operations	data big platforms circles insight storage managing analysis models	1.45
Life science	health analytics life care ideas impact medical science process	1.44
Technical support	solution customer technical support creativity task meeting role effort	1.22
Total		**36**

developed for big data. The taxonomy consisted of five competency domains that are: domain-specific competencies (36%), developer competencies (21%), soft competencies (18%), business-oriented competencies (14%), and analytical competencies (11%). The percentages given with competency domains indicate the total percentage of topics' rate in the related field. In the following subsection, the competency domains are given under five distinct subheadings in order to provide a better understanding of the results.

3.1 Domain-specific competencies

Of all the discovered topics, 20 were related to domain-specific knowledge and skills. The sum of the percentages of these topics is 36%. As seen in Table 1, the collection of domain-specific knowledge and skills cover considerable educational background (2.57%), and experience (2.56%). Besides, technical skills such as big data platforms, database technologies, and streaming data processing are highly demanded. On the most abstract level, domain-specific competencies consist of deep-domain knowledge, and up-to-date big data tools, technologies, and frameworks.

3.2 Developer competencies

Of all the discovered topics, 12 were related to developer competencies. The sum of the percentages of these topics is 21%. Given the findings related to developer competencies in Table 2, the topics highlight the current tendencies for development processes of big data applications. Especially, the topic of "script programming" has the highest rate (2.62%) in all of them. This finding indicates the significance of script programming for big data applications. Likewise, the topics of web services, agile development, cloud computing, devops automation, and distributed systems are other important tendencies with high rates. In a general perspective,

Table 2. Topics related to developer competencies

Topic Label	Descriptive Keywords	Rate %
Script programming	java python programming language scala scripting shell spring	2.62
Web services	web api service rest set connect restful app aws pi	2.06
Agile development	development software agile scrum methodology process lifecycle	1.97
Cloud computing	cloud aws services amazon computing emr azure google redshift	1.93
Devops automation	tools continuous automation devops integration jenkins deployment build	1.80
Distributed systems	system large scale distributed scalable complex multi computing designing	1.76
Development processes	code practice process development standards software testing application design	1.75
Object-oriented	design architecture implementation oriented development object practice	1.72
Testing	developing test testing production implementing quality designing cases unit	1.67
Performance tuning	performance application system tuning monitoring support scalability debugging	1.58
Back-end development	application end back infrastructure user building components developing	1.46
Software management	product software management manager responsible features development	1.34
Total		**21**

the conceptual framework of the developer competencies covers processes, methodologies, and platforms used in application development for big data.

3.3 Soft competencies

Of all the discovered topics, 11 were related to soft competencies. The sum of the percentages of these topics is 18%. The topics are presented in Table 3 with their rates. According to the findings, the topic of "communication skills" has a high rate (2.58%) among the topics. This means that the topic of communication skills was one of the indispensable qualifications for big data professionals. Likewise, project management, computational thinking, learning skills, collaboration and leadership are among the highly demanded soft competencies. Big data life-cycles require an active interaction of many actors in different roles. For this reason, the soft competencies are based on high-level communication, collaboration, and cognitive skills.

3.4 Business-oriented competencies

Of all the discovered topics, 10 were related to business-oriented competencies. The sum of the percentages of these topics is 14%. As seen in Table 4, the collection of business-oriented competencies contains specific knowledge domains for business such as business processes, business management, human resource, business solutions, and sales-marketing and decision-making competencies. The findings in the table also indicated the necessity of business-oriented competencies in order for big data operations to be effectively used in decision-making processes of businesses.

Table 3. Topics related to soft competencies

Topic Label	Descriptive Keywords	Rate %
Communication skills	skills ability communication problem written solving strong excellent verbal	2.58
Project management	project multiple manage tasks planning needed guidance priorities timely	1.79
Computational thinking	computational critical thinking innovation strategy making effect	1.77
Learning skills	setting work experience fast paced learn apply dynamic quickly	1.72
Collaboration	team work lead members engineers independently collaborate member closely	1.62
Leadership	leadership support strategic role strategy activities goals direction initiatives	1.61
Technical communication	business technical internal communicate effectively external reports stakeholders	1.52
Problem solving	problems complex issues solution provide solve process identify critical analytical	1.39
Creativity	technology creativity vision build global mobile create innovative building	1.34
Team working	team join edge engineer seeking cutting opportunity working role part	1.28
Self-development	self work career development individual environment build growth clients	1.27
Total		**18**

Table 4. Topics related to business-oriented competencies

Topic Label	Descriptive Keywords	Rate %
Business processes	requirements technical business functional processes existing meet define	1.58
Business management	management IT business industry expertise knowledge deep provide qualified	1.48
Human resource	opportunity employment resource human manage equal apply contact	1.44
Business solutions	solution technology business delivery working functional lead cross global	1.44
Sales & marketing	business solution develop sales key market customer strategy drive relationships	1.43
Customer services	customer team service solution organization wide product success drive variety	1.40
Decision-making	data business work decision making core key judgment successful system	1.36
Career development	company opportunities employees career world grow create growing helping	1.33
Business mission	services leading mission operations technology world company customer critical	1.33
Digital marketing	digital platform marketing creating deliver clients insight management consulting	1.32
Total		**14**

Table 5. Topics related to analytical competencies

Topic Label	Descriptive Keywords	Rate %
Data reporting	tools etl reporting BI modeling visualization integration informatica tableau	1.87
Machine learning	learning machine techniques including models methods principles concepts	1.67
Business intelligence	business intelligence descriptive predict analysis model estimate	1.65
Advanced analytics	data analytics big advanced solution analytic intelligence responsible leverage	1.43
Research analysis	analysis research quality support develop data improvement intelligence trends	1.36
Social networks	social media data analytics mining live tendency community twitter	1.27
Financial analytics	services financial analytics finance banking corporate leading prediction	1.27
Total		**11**

3.5 Analytical competencies

Of all the discovered topics, seven were related to analytical competencies. The sum of the percentages of these topics is 11%. The findings as given in Table 5, indicated that the analytical competencies underline analytical methods and methodologies consisting of data reporting, machine learning, business intelligence, advanced analytics, research analysis, and so on. Also, the methods and methodologies require a scientific background based on analytics, mathematics, and statistics.

4. Discussion

The findings of this study revealed the core competencies (knowledge, skills, and abilities) required for big data in detail. Initially, the findings indicate that expertise of big data requires an interdisciplinary background and a collection of competency domains including domain-specific competencies, developer competencies, soft competencies, business-oriented competencies, and analytical competencies. Our findings are also consistent with the results of earlier studies and industry reports highlighting the necessity of a variety of competency sets to process, analyze, and manage big data in improving business processes [5, 9, 18]. Especially, the necessity of communication skills as well as technical skills and analytical skills has been also stated in the previous studies [4, 9, 18]. The discovered topics by LDA also uncover the emerging trends and technologies (e.g., platforms, programming languages and databases) in big data systems as well as the competency requirements for big data professionals. Big data platforms such as Hadoop, Spark, and Hive, programming languages such as Java and Python, and databases such as SQL and NoSQL are also reported as leading technologies that maintain their dominance in big data field [5, 9, 21, 30]. Besides, our findings demonstrate that core competencies encompass a wide spectrum of knowledge domains and skill sets which implies that big data professionals can undertake different roles and responsibilities in different big data workflows. Depending on the competencies achieved at the end of an effective CBEE period, the professional can be employed in different positions with different job titles. For instance, at the end of a CBEE period, candidates who achieve analytical competencies can work as business analysts or data analysts. Likewise, candidates who achieve developer competencies can work as big data developers [18, 19].

As a result, the findings reveal the necessity of an innovative and CBEE-focused big data curriculum based on interdisciplinary collaboration for related programs. CBEE is a demand-driven approach and the first priority in this model is to focus on industrial needs and demands, and implement an education process that can meet those demands. Especially in terms of bridging the gap between academia and engineering industries, the necessity of CBEE-focused programs has been addressed by earlier studies as well [1, 6, 11, 21]. In order to provide an effective CBEE program, the standards and requirements demanded by the industry must be clearly addressed in these programs. In this regard, the findings of this study provide a significant implication for the CBEE, especially in academic preparations before the education program. Our findings constitute an important source of information for academic preparations before the CBEE program, such as the selection of competencies to be included in the education program, designing and updating of course contents, and determination of the learning objectives [18, 19]. Considering the CBEE for big data, core competencies can be categorized as basic, common and job-specific competencies. The domain-specific competencies can be considered as basic-level core courses. The soft competencies can be considered as common courses. Likewise, the developer competencies, analytical competencies, and business-oriented competencies can be considered as job-specific courses. The distribution of the time and the credits of these courses can be organized according to the topics rates. Consequently, the CBEE-oriented methodology recommended for big data can also be applied to dynamic engineering disciplines such as robotics, cloud computing and software engineering [12, 13, 18–21].

5. Conclusion

This paper attempts to shed light on the core competencies for big data and to obtain implications that may set targets for CBEE programs. With this aim, a topic-modeling based content analysis was performed on big data job ads. The methodology of the study is based on LDA, a hierarchical Bayesian approach for probabilistic topic modeling widely-applied in text mining, which discovers trending topics from the textual content of the job ads. As a result of this analysis performed by LDA-based topic modeling, the 60 trending topics reflecting essential knowledge domains and skill sets for big data were discovered in a finer granularity level. By analyzing and interpreting these topics, we developed a competency taxonomy for big data containing domain-specific competencies, developer competencies, soft competencies, business-oriented competencies, and analytical competencies. Our research provides valuable insights for better understanding of the main characteristics and tendencies of big data jobs. The findings of the study have meaningful implications for the main big data actors from different aspects. At the institutional level, the findings may help companies to identify qualified big data professionals, and they may help academic institutions to meet the need for a qualified big data workforce. At the individual level, the findings may be beneficial for big data professionals in assessing and updating their own skills, for instructors in educating big data candidates in line with emerging demands, and for students in giving direction to their careers. Additionally, our research methodology can be applied to other engineering disciplines.

References

1. J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh and A. H. Byers, *Big data: The next frontier for innovation, competition, and productivity*, 2011.
2. F. Provost and T. Fawcett, Data Science and its Relationship to Big Data and Data-Driven Decision Making, *Big Data*, **1**, pp. 51–59, 2013.
3. M. Minelli, M. Chambers and A. Dhiraj, *Big data, big analytics: emerging business intelligence and analytic trends for today's businesses*, John Wiley & Sons, 2012.
4. H. Chen, R. H. L. Chiang, V. C. Storey, Business Intelligence and Analytics: From Big Data to Big Impact, *Mis Quarterly*, **36**, pp. 1165–1188, 2012.
5. S. Debortoli, O. Müller and J. vom Brocke, Comparing Business Intelligence and Big Data Skills, *Business & Information Systems Engineering*, **6**, pp. 289–300, 2014.
6. S. Miller, Collaborative Approaches Needed to Close the Big Data Skills Gap, *Journal of Organization Design*, **3**, p. 26, 2014.
7. D. Stevens, M. Totaro and Z. Zhu, Assessing IT critical skills and revising the MIS curriculum, *Journal of Computer Information Systems*, **51**, pp. 85–95, 2011.
8. A. De Mauro, M. Greco, M. Grimaldi and P. Ritala, Human resources for Big Data professions: A systematic classification of job roles and required skill sets, *Information Processing and Management*, **54**(5), pp. 807–817, 2017.
9. A. Gardiner, C. Aasheim, P. Rutner and S. Williams, Skill Requirements in Big Data: A Content Analysis of Job Advertisements, *Journal of Computer Information Systems*, **58**(4), pp. 374–384, 2017.
10. P. Russom, Big data analytics, *TDWI Best Practices Report, Fourth Quarter*, **19**(4), pp. 1–34, 2011.
11. R. Dubey and A. Gunasekaran, Education and training for successful career in Big Data and Business Analytics, *Industrial and Commercial Training*, **47**(4), pp. 174–181, 2015.
12. R. Kelchen, The landscape of competency-based education: Enrollments, demographics, and affordability, *American Enterprise Institute for Public Policy Research*, **21**, 2015.
13. V. Kovaichelvan, Competency-based engineering education, *International Journal of Indian Culture and Business Management*. **8**(2), pp. 253–273, 2014.
14. S. Sheppard, A. Colby, K. Macatangay and W. Sullivan, What is engineering practice?, *International Journal of Engineering Education*. **22**(3), p. 429, 2007.
15. J. R. Goldberg, V. Cariapa, G. Corliss and K. Kaiser, Benefits of industry involvement in multidisciplinary capstone design courses, *International Journal of Engineering Education*, 2014.
16. M. Daniels, Å. Cajander, T. Clear and R. McDermott, Collaborative technologies in global engineering: new competencies and challenges, *International Journal of Engineering Education*, **31**(1), pp. 267–281, 2015.
17. T. J. Brumm, L. F. Hanneman and S. K. Mickelson, Assessing and developing program outcomes through workplace competencies, *International Journal of Engineering Education*, **22**(1), p. 123, 2006.
18. S. Mamonov, R. Misra and R. Jain, Business analytics in practice and in education: A competency-based perspective, *Information Systems Education Journal*, **13**(1), p. 4, 2015.
19. H. Topi, IS EDUCATION Using competency-based approach as foundation for information systems curricula: benefits and challenges, *ACM Inroads*, **7**(3), pp. 27–28, 2016.
20. P. Boahin and W. H. A. Hofman, Perceived effects of competency-based training on the acquisition of professional skills, *International Journal of Educational Development* **36**, pp. 81–89, 2014.
21. F. Gurcan and C. Kose, Analysis of Software Engineering Industry Needs and Trends: Implications for Education, *International Journal of Engineering Education*, **33**(4), pp. 1361–1368, 2017.
22. D. M. Blei, A. Y. Ng and M. I. Jordan, Latent Dirichlet Allocation, *Journal of Machine Learning Research*, **3** pp. 993–1022, 2003.
23. D. M. Blei, Probabilistic topic models, *Communications of the ACM*, **55**(4), pp. 77–84, 2012.
24. Job Search | Indeed, (n.d.), https://www.indeed.com, Accessed 26 March, 2018.
25. A. Karl, J. Wisnowski and W. H. Rushing, A practical guide to text mining with topic extraction, *Wiley Interdisciplinary Reviews: Computational Statistics*, **7**(5), pp. 326–340, 2015.
26. A. K. Uysal and S. Gunal, The impact of preprocessing on text classification, *Information Processing & Management*, **50**(1), pp. 104–112, 2014.
27. T. L. Griffiths and M. Steyvers, Finding scientific topics, *Proceedings of the National Academy of Sciences*. **101**(1), pp. 5228–5235, 2004.
28. A. K. McCallum, MALLET: A Machine Learning for Language Toolkit, http://mallet.cs.umass.edu, Accessed 22 May 2018.
29. H. M. Wallach, Topic Modeling: Beyond Bag-of-Words, *In Proceedings of the 23rd international conference on Machine learning*, pp. 977–984, 2006.
30. F. Gürcan and M. Berigel, Real-Time Processing of Big Data Streams: Lifecycle, Tools, Tasks, and Challenges, *In 2018 2nd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*, pp. 1–6, 2018.

Fatih Gurcan, PhD, is an instructor in Department of Informatics. He received BS degree in Department of Statistics and Computer Sciences, and MS degree in Department of Computer Engineering from Karadeniz Technical University, Trabzon, Turkey. He received the PhD degree in the Department of Computer Engineering from Karadeniz Technical University, Trabzon, Turkey. His research interests contain big data, engineering education, competency-based education, sentiment analysis, topic models, big data analytics, and text mining.