Service Providers' Competence Identification in Knowledge-Intensive Crowdsourcing Context

Biyu Yang
School of Mechanical Engineering
Chongqing University
Chongqing, China
yangby123@outlook.com

Xu Wang
School of Mechanical Engineering
Chongqing University
Chongqing, China
wx921@163.com

Quanwu Zhao
School of Business and Economics
Chongqing University
Chongqing, China
zhaoquanwumx@cqu.edu.cn

Zhuofei Ding
School of Informatics
University of Edinburgh
Edinburgh, UK
dingzhuofei@163.com

Mei Long
Administrative Department
ZBJ Co. Ltd
Chongqing, China
longmei@zbj.com

2020 9th International Congress on Advanced Applied Informatics (IIAI-AAI) | 978-1-7281-7397-9/20/$31.00 ©2020 IEEE | DOI: 10.1109/IIAI-AAI50415.2020.00027

***Abstract*—Competence analysis of service providers is of great importance for a knowledge-intensive crowdsourcing platform to examine service providers' (SPs) effectiveness and contribution and therefore to improve its management and operation efficiency and accuracy. As the information highway highly-developed, there are a huge amount of online crowdsourcing communities where talent workers can exchange experience and ideas, which enables competence analysis using text mining. In this paper, we developed a competence identification framework to analyze and recognize SPs' competence in online knowledge-intensive crowdsourcing (KIC) context. We firstly crawled the experience sharing articles, which contained excellent SPs' experiences and ideas about the efforts that should be made to be successful, from several Chinese online crowdsourcing communities and then text mining techniques were applied to analyze these unstructured texts. Each sentence in the corpora was tokenized into several words, after which the words were clustered as different topics using Latent Dirichlet Allocation (LDA) model based on their underlying semantics. Furthermore, based on the LDA outputs, we identified six clusters of crowdsourcing SPs' competence and thus constructed the competence system on the basis of Spencer's competence dictionary and human intervention. Finally, the descriptions of the competence system were presented.**

Keywords—competence analysis; crowdsourcing; innovation design; LDA; service provider evaluation

I. INTRODUCTION

Crowdsourcing has become a powerful mechanism for accomplishing work online[1]. Early successful cases of crowdsourcing including Wikipedia, Yelp Yahoo! Answers and Amazon Mechanical Turks (AMT)[2], have made crowdsourcing more powerful and manageable. At the same time, the increased use of the Internet and web-based crowdsourcing platforms, has provided many advantages because they support and enhance the connection between service providers with their customers[3].

Knowledge-intensive crowdsourcing platforms (KI-CPs) tap into the creative fields of intelligent crowd (e.g., designers), changing the traditional way how business is conducted[4]. For example, in the design industry, penetration of technology, especially internet and smart phone apps, has changed how design is practiced, produced, assessed, traded, taught and learned. Quantities of service providers (SPs), with a variety of skills and expertise, gather on the KI-CPs and can be accessed easily by varied service demanders (SDs). In general, a service demander (SD) will post a task statement outlining the details of task, a proposed price, and a deadline at the crowdsourcing platform. Then a service provider (SP) interested in the task will positively communicate with the SD or bid for the task directly. The SD chooses one or several SPs to perform the task and rewards them according to the quality of their service products. It is an obvious fact that the products' quality surely depends on SPs' competence. If a SP has a high level of capability and competence, SDs' requirements can be possibly fulfilled and final service products by the SP can be of high quality. Consequently, SDs will be satisfied, willing to continuously interact with the SP and the platform. However, a SD will not know whether the SP is appropriate for the task until they actually receive the final products or outputs[5]. Moreover, the number of SPs increases rapidly each year, on the one hand, competition is getting fierce and SPs should have self-knowledge and recognize the way to survive and sustainably develop. On the other hand, it is harder and difficult for KI-CPs to meet SPs' needs systematically. If KI-CPs can understand and master SPs' competence clearly, then a full range of value-added services can be designed and offered to facilitate SPs' competence growth, thus improving the sustainable and competitive development of the platform.

In this paper, we developed a competence identification framework to analyze and recognize service providers' competence in online knowledge-intensive crowdsourcing (KIC) context for distinguishing their effectiveness and contribution, and therefore control the quality of KIC outcomes. We applied several text mining techniques to unstructured text materials obtained from several Chinese online crowdsourcing communities to extract SPs competence features. Each sentence in the unstructured text was separated into several words. After the general text preprocessing, we inputted the words set into Latent Dirichlet Allocation (LDA) model. We analyzed the LDA outputs and clustered six

This work is supported by the National Science and Technology Support Program of China (Project NO. 2018YFB1403600 and NO. 2018YFB1403602), the Technological Innovation and Application Program of Chongqing (Project No. cstc2018jszx-cyzdX0081), and the Fundamental Research Funds for the Central Universities (Project NO. 2020CDCGJX019).

978-1-7281-7397-9/20/$31.00 ©2020 IEEE
DOI 10.1109/IIAI-AAI50415.2020.00027

competence dimensions on the basis of competence theory. Furthermore, we constructed a competence system of crowdsourcing SPs working in innovative design industry.

We begin this paper with a background introduction of crowdsourcing and competence analysis. Next, a literature review of relevant research is presented in Section 2. The data collection and text mining techniques of our research are detailed in Section 3, followed by the results and descriptions of the competence system in Section 4. Finally, a conclusion is drawn in Section 5.

II. Literature Review

Several research fields in operations management, crowdsourcing, and machine learning are relevant to our work on service providers' competence analysis in big-data settings. These include the literatures on competency theory, SPs' evaluation in crowdsourcing environment and the natural language processing (NLP) working on topic models in big data context. We briefly review these fields below.

A. Competence

The concept of competence was initially introduced by McClelland in his paper "testing for competence rather than intelligence"[6] in the 1970s. As the competence theory can provide significant help with key problems, for example, how to clarify workforce standards and expectations, and how to align individual employees with the organization's business strategies[7], it has been widely applied, moving from the field of human resource management[8, 9] to various business disciplines[10]. Generally, competence are characteristics of people that differentiate performance in a specific job or role[6]. In the book "Competence at work: Model for superior performance"[11], Spencer has proposed a competence dictionary and has constructed five generic competence models for technicians, salespeople, helping and human service professionals, managers and entrepreneurs, respectively. In his work, Spencer concluded six clusters of competence regarding to achievement and action, helping and human service, the impact and influence, managerial, cognitive, and personal effectiveness, and 21 scales of competence dimensions. The generic dictionary scales are applicable to all jobs[11] and we referred to the dictionary to develop the competence system for SPs in KIC environment in this paper.

B. SPs' Evaluation in Crowdsourcing Context

In crowdsourcing context, SPs are mainly evaluated based on their expertise and reliability for a work or an activity for pre-selecting persons. In general, pre-selection is defined as "a means of ensuring a minimum ex-ante quality level of contributions"[12], which is thought to be a proper way to mitigate the risk of poor quality solutions by screening potential SPs based on the completion of some process that indicates certain knowledge or expertise[13, 14].

In the context of crowdsourcing, there are several evaluation means and techniques provided by platforms to pre-select a SP. For instance, CrowdFlower applies a so-called gold standard strategy where a strategy by some items are injected into the annotations with known labels to evaluate the annotators and thus eliminating the spammers[15]. According to [16], in their work of assessing Text Retrieval Conference (TREC) relevance using crowdsourcing, both the qualification test and golden standard method were used to compare the accuracy between Mechanical Turk (MTurk) workers and TERC assessors. [17] concluded the factors that affect the reliability of crowdsourcing outputs, including experimental design[18], human features and monetary factors, and explored how SPs' verbal comprehension skill affected their competence by a human gold standard in the context of relevance assessments in the evaluation of information retrieval systems. [19] proposed a combinatorial optimization algorithm to select a group of SPs based on their reliability estimated by a set of gold standard questions with known answers.

Although these evaluation methods mentioned above are simple to apply and typically perform well, there still are some shortcomings. On the one hand, pre-selection techniques may be useful to prevent "scammer" SPs to participate in, but some industrious SPs may not select the task because of extra burden caused by human involvement[13, 20]. On the other hand, answers to these test may be shared among applying SPs, which leads to the effectiveness reduction.

Besides, other algorithm-based methods are also applied to analyze and measure SPs' expertise and skills. [15] presented an empirical Bayesian algorithm to iteratively eliminate the spammers and evaluate the consensus labels based only on the good annotators. [21] proposed a Bayesian generative probabilistic model for the annotation process, to identify the image difficulty, annotator competence and bias. [20] proposed a general worker quality evaluation algorithm without using pre-developed answers and then applied the proposed algorithm in the Hadoop platform using MapReduce programming model for parallel evaluation for a loads of SPs in big data context. In order to identify experts in crowdsourcing engineering design evaluation, [22] compared four expertise prediction heuristics, namely evaluation demographics, evaluation reaction times, mechanical reasoning aptitude and easy version of evaluation task.

However, these evaluation methods are task-oriented [23], and can't insightful details of SPs capabilities and competence, with the outputs of these methods only present whether a specific SP is allowed to perform a specific task. Therefore, there is still scant research that explore SPs' competence in KIC environment in a more comprehensive and specific way.

C. LDA Model

The rising development and accessibility of large electronic archives, along with increased computational facilities, has led to the interest in textual content analysis [24-26].

Topic modelling is effective in extracting implied themes in large-scale text based on word co-occurrence for each document in the corpus. What's more, topic modeling is an useful way to "let the text talk" due to the independent on the evaluator's personal opinions or experiences[27, 28]. LDA model is a well-known unsupervised machine learning techniques for natural language processing[29] and is the simplest and most popular topic modeling algorithm[30] to be used to recognize the hidden topics and mine deep semantics of textual documents. [31] proposed a LDA model to analyze consumer complaints of the Consumer Financial Protection Bureau (CFPB), aiming to extract latent topics in the complaint narratives and explore their associated trends over time. [30] applied LDA model to E-petitions analysis and as a result 87% of generated topics were meaningful to human judges. [32] identified job satisfaction factors by applying LDA to online employee reviews. The sentiment and importance of each job satisfaction factor were measured at industry, company, group, and chronological levels. To identify the customers' views and opinions regarding online retail brands, [33] used a combination of text analytical approaches including LDA, sentiment analysis, and network analysis to analyze the tweets associated with five leading UK online retailers. To improve customer service of cross-border e-commerce, [34] employed LDA to explore complaint topics from consumer feedbacks on e-commerce platforms. [35] excavated user-topics distribution based on a Hashtag-LDA model to recommend relevant microblogs to users. In the domain of Web service, [29] proposed a new domain-aware Mashup service clustering method by leveraging LDA topic model based on multiple data sources and enhanced the results of Mashup service discovery.

Since its popularity and effectiveness in topic modelling, in this paper, we use LDA model to extract SPs competence leveraging the data from crowdsourcing communities.

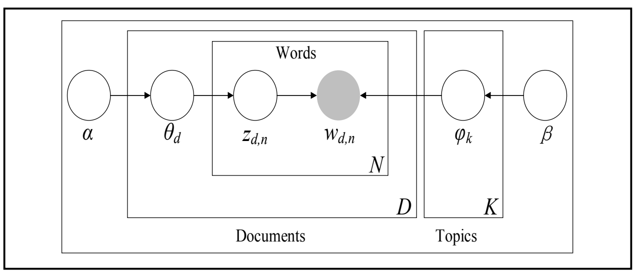

Fig. 1. The generative probabilistic graph model of LDA

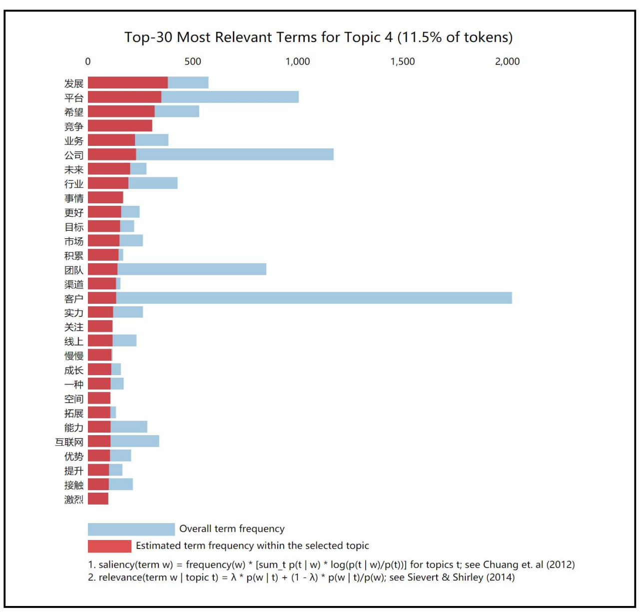

Fig. 2. An example of a LDA topic

III. Competence Identification Framework

In this part, we give details of our competence identification framework on data collection preprocessing and the LDA model for extracting SPs' competence in KIC context. These procedures were conducted on an Intel core i5 CPU, 12 GB RAM machine. The raw data were crawled and the LDA model was constructed based on Scikit-learn under Python version 3.6.5.

A. Data Collection

In China, there are several popular crowdsourcing communities that will invite high-performance SPs to share their experience regarding to performing more tasks and getting more money. This will be conducted in the form of interviews, that those SPs need to answer a series of questions involving different aspects of being successful. Then, the interview records are organized as series of articles posted online for other SPs to learn from. We think these experience sharing articles are of great significance for platforms to derive SPs effective competency factors that can differentiate their performance. To that end, we developed a crawl program coded in Python, and collected experience sharing articles in innovation design sector as our corpora by crawling. Information including title, poster, post time, view times, comments and responses were obtained. Finally, 1760 raw articles from 14/05/2018 to 19/07/2019 were collected.

B. Data Preprocessing

The raw articles collected in online crowdsourcing communities are all unstructured and in Chinese. We applied the general Chinese text preprocessing steps to clean the unstructured text for topic modeling. To start with, we eliminated all numbers and alphabets, just leaving the Chinese characters. Then, each article was cut into sentences which were next tokenized into multiple space-separated words. To better remove the stop words, except for using the popular stop-words list from Harbin Institute of Technology [36], we constructed our own stop-words list where these words are design-specific and highly appear in our corpora but useless for analysis. After removing stop words, we obtained the cleaned words set.

C. Competence Extraction

We use LDA to grasp the underlying structures of experience-sharing articles. LDA model introduces Dirichlet prior and offers a fully generative model, in which each document is represented as a multinomial distribution of topics. These topics can be regarded as higher level concepts that are akin to clusters. There is a basic assumption that each document in a database is generated from several latent topics, where each topic is presented by a mixture of words. The generative probabilistic model of LDA is represented in Fig 1. In the LDA model, K is the number of topics; D, the number of articles; N_d, the length of the d_{th} article; θ_d, the ability distribution of the d_{th} article; φ_k, the word probability distribution of the topic k; $w_{d,n}$, the n_{th} word in the d_{th} article; and $z_{d,n}$, the word's topic. Parameters α and β are the hyperparameters for prior distributions of θ_d and φ_k, respectively. Among these variables, $w_{d,n}$ are observable and the number of topics K are exogenous given. Given the number of topics, LDA assumes the following generative process for each article d in a corpus D.

1)For each experience sharing article d, select a multinomial θ_d that is originated from Dirichlet hyper-parameter α.

2)For each topic k, select a multinomial φ_k that is originated from Dirichlet hyper-parameter β.

3)For each word $w_{d,n}$ in the article d, perform below:

(a)Choose a topic $z_{d,n}$, which belongs to Multinomial $\theta_{d,n}$ distribution.

(b)Choose a word $w_{d,n}$ from $p(w_{d,n} \mid z_{d,n}, \beta)$ (i.e., $\varphi_{d,n}$), where $\varphi_{d,n}$ follows a multinomial distribution.

The outputs of LDA are per-document topic distributions $z_{d,n}$ and per-word topic distributions $w_{d,n}$. In this work, as our purpose is to identify competence features of SPs, the words within each topic and their underlying meaning have high significance. According to experts' opinion, the number of topics K was set to 10. We further illustrate the generated topics using pyLDAvis in Python. Fig.2 shows an example of one LDA topic.

IV. RESULT

The purpose of this research is to analyze and identify important competence features of SPs in KIC context. However, different industry fields require different and unique competence features. To be more targeted, we collect textual data related to innovation design industry from several crowdsourcing communities in China. Several text analysis methods, including general text preprocessing and LDA model to clean the raw material and extract hidden topics from the cleaned words set according to their frequency and semantics. Referring to Spencer's generic competence dictionary[11], we mapped the LDA topics into six clusters, namely design ability, innovation ability, sociability, service awareness, competitive spirit and team composition, to differentiate SPs capability and performance. Fig. 2 shows an example of one of the LDA topics and Table I presents the competence system.

To sum up, in the context of innovation design, SPs working on innovation design tasks should develop and obtain their competence from the six aforementioned aspects. The description of each competence cluster is stated as follows.

(1) Design Ability. As innovation design is the focus in our research, it's no doubt that the related SPs should possess design related abilities, which is the first and most important point that a customer deeply values. The professional skills, related experience and the dedication will enhance the likelihood of high-quality outcomes and successful transactions. Therefore, we concluded this topic as design ability.

(2) Innovation Ability. As to innovation design, it is the creative ideas that make customers accept the final products at most of the time. Also, local culture is of great significance in design industry. Specifically, Chinese customers may prefer designs which are integrated into traditional Chinese cultures. Besides, the underlying concepts behind a work are also appealing and valued. Therefore, we concluded this topic as innovation ability.

(3) Sociability. It is much harder for both SDs and SPs to trust into each other in online markets than offline ones. However, this can be mitigated by reciprocal communication and information exchange. If a SP has high-level communication skills and can solve problems with customers together, mutual trust may be constructed quickly, then the tasks will be better completed. Therefore, we concluded this topic as sociability.

(4) Service awareness. SPs should take initiative when interacting with customers. In the context of online crowdsourcing, both sides are connected by a specific task. To make customers' satisfied, SPs should focus on the task, solve customer's problems and meet their needs positively. For example, to finish a package design, a customer may require revisions for several times. During the whole serving process, the SP should communicate with the customer actively, understand and respond to customer's requirements accurately. Therefore, we concluded this topic as service awareness.

(5) Competitive Spirit. There are thousands of tasks waiting to be selected and matched in the system, and also quantities of SPs competes in the same field. However, the time, energy, and labor forces are limited. It is wise for SPs to make a choice according to the reward and complexity of a task, and take their own abilities into account at the same time. The procedure of completing a task is also a process of learning, summary, self-development and competition edge obtaining. SPs should make efforts to stand out from the mass peers. Therefore, we concluded this topic as competitive spirit.

(6) Team Composition. In China, most online crowdsourcing SPs work as a team, within which the members, with functional heterogeneity and have different backgrounds and knowledge, will complement the way to obtain information benefits and business advantages. Meanwhile, a good team can encourage and collaborate with each other and make progress in an alarming rate. Therefore, we concluded this topic as team composition.

V. CONCLUSION

In this paper, we propose a competence identification framework to analyze and recognize effective competences of SPs working in KIC environment. To start with, we collected experience sharing articles regarding to the efforts that should be made to be a successful SP in innovation design industry in several Chinese crowdsourcing communities. Then, we applied the general Chinese text preprocessing procedures to the initial text, during which, we constructed our own design-related stop-words list. Further, the LDA model were applied to the cleaned words set at the next step. Finally, referring to Spencer's generic competence dictionary, we concluded the

LDA topics into six competence clusters and established the competence system of SPs in innovation design industry.

TABLE I. THE COMPETENCE SYSTEM OF INNOVATION DESIGN

Competence Clusters	Competence Features
Design Ability	设计 (design)、策划 (plan)、订单 (orders)、中标 (bid)、技术 (expertise)、专业 (professional)、服务 (service)、经验 (experience)、交易 (transaction)、时间 (time)、提供 (provide)、系统 (systematic)、产品 (product)、文案 (copy)、研发 (research & development)、定制 (customization)、参加 (participation)、工作 (work)、完善 (improvement)、成功 (success)、评价 (comment)、沟通 (communication)、努力 (diligence)、交流 (contact)
Innovation Ability	创意 (originality)、文化 (culture)、中国 (China/Chinese)、价值 (value)、专业 (professional)、理念 (idea)、创新 (creativity)、产品 (product)、宣传 (advertisement)、作品 (production)、创造 (creation)、标志 (logo)
Sociability	朋友 (friend)、在线 (online)、机会 (opportunity)、印象 (impression)、理解 (understand)、角度 (perspective)、接受 (accept)、信息 (information)、世界 (world)、想到 (taking into consideration)、真诚 (sincere)、深刻 (profound)、合作伙伴 (partner)、思考 (thinking)、互动 (interaction)、关键 (key)
Service Awareness	客户 (customer)、沟通 (communication)、设计 (design)、服务 (service)、作品 (production)、需求 (request)、满意 (satisfaction)、修改 (modification)、用心 (concentration and diligence)、态度 (altitude)、做好 (improve)、过程 (process)、建议 (suggestion)、耐心 (patience)、提出 (suggest)、负责 (responsible)、诚信 (loyalty)、方案 (plan)、效果 (effect)
Competitive Spirit	发展(development)、希望 (vision)、竞争 (compete)、业务 (business)、未来 (future)、梦想 (dream)、行业 (industry)、更好 (improve)、目标 (business goal)、学习 (learning)、积累 (accumulation)、经历 (experience)、知识 (knowledge)、渠道 (channel)、实力 (ability)、线上 (online)、关注 (focus)、事业 (career)、成长 (development)、能力 (capability)、空间 (potential)、拓展 (expansion)、优势 (advantage)、提升 (promotion)、选择 (choice)、人生 (lifetime)、时代 (era)
Team Composition	参与 (involvement)、团队 (team)、喜欢 (enjoy)、成员 (members)、信心 (faith)、理想 (dream)、一群 (a group)、充满 (full of)、鼓励 (encourage)、热爱 (enthusiasm)、激情 (passion)、动力 (motivation)

The competence identification framework contributes to the following aspects: (1) both platforms and SPs should pay attention to the six competence clusters and these keywords, for platforms can establish and construct working standards for SPs to guarantee the quality of their services and products, and for SPs to make self-check and self-improvement by realizing their advantages and disadvantages, (2) a competence measure instruction can be constructed by crowdsourcing platforms to screen, select, manage SPs to allocate customer and order resources strategically, and (3) the platform can develop value-added service products to assist SPs in improving their capabilities and performance.

In addition, there are some limitations in our study. First, although the topic modelling approach (e.g., LDA) shows popular applications in text analysis, it still has some shortcomings. As an unsupervised learning algorithm, LDA inherently has disadvantages in fully understanding natural languages but it requires no human intervention. Future research may use supervised learning algorithms for identifying SPs competencies from online experience sharing data. Second, our research is limited by the single source of materials, which may result in incomplete competence identification. It might be necessary to collect and use more data from different data sources in future research to obtain more generalized and significant findings. Finally, it may be worthwhile to extend our study to that it could include other industries, other text mining techniques, and other data as future research for more significant analytical results on SPs' competence in KIC context.

ACKNOWLEDGMENT

The authors would like to thank the members in the team of R&D of technology consulting service platform for their assistance and guidance. The authors also appreciate the editor and the anonymous referees for their manuscript processing and remarkable comments.

REFERENCES

[1] Kittur, A., B. Smus, S. Khamkar and R.E. Kraut, "Crowdforge: Crowdsourcing complex work," in UIST'11 - Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology. 2011.

[2] Kittur, A., E.H. Chi and B. Suh, "Crowdsourcing user studies with mechanical turk," Conference on Human Factors in Computing Systems - Proceedings, 2008, pp. 453-456.

[3] Kitsios, F., S. Stefanakakis, M. Kamariotou and L. Dermentzoglou, "E-service evaluation: User satisfaction measurement and implications in health sector," Computer Standards and Interfaces, 2019, 63, pp. 16-26.

[4] Tian, X., J. Shi and X. Qi, "Talent crowdsourcing via stochastic sequential assignments," SSRN Electronic Journal, 2018.

[5] Igawa, K., K. Higa and T. Takamiya, "An exploratory study on estimating the ability of high skilled crowd workers," Proceedings - 2016 5th IIAI International Congress on Advanced Applied Informatics, IIAI-AAI 2016, 2016, pp. 735-740.

[6] McClelland, D.C., "Testing for competence rather than for "intelligence"," The American psychologist, 1973, 28, (1), pp. 1-14.

[7] Wu, W.W. and Y.T. Lee, "Developing global managers' competencies using the fuzzy dematel method," Expert Systems with Applications, 2007, 32, (2), pp. 499-507.

[8] Ploum, L., V. Blok, T. Lans and O. Omta, "Toward a validated competence framework for sustainable entrepreneurship," Organization and Environment, 2018, 31, (2), pp. 113-132.

[9] Siswanti, D.N., R. Khairuddin and F. Halim, "The effect of spiritual intelligence, emotion and social competence to the leadership competence," Journal of Physics: Conference Series, 2018, 1028, (1).

[10] Barnes, J. and Y. Liao, "The effect of individual, network, and collaborative competencies on the supply chain management system," International Journal of Production Economics, 2012, 140, (2), pp. 888-899.

[11] Spencer, L.M. and S.M. Spencer, Competence at work: Models for superior performance. 1993: John Wiley & Sons.

[12] Geiger, D., S. Seedorf, T. Schulze, R. Nickerson and M. Schader, "Managing the crowd: Towards a taxonomy of crowdsourcing processes," 17th Americas Conference on Information Systems 2011, AMCIS 2011, 2011, 5, pp. 3796-3806.

[13] Dukat, C. and S. Caton, "Towards the competence of crowdsourcees: Literature-based considerations on the problem of assessing crowdsourcees' qualities," in IEEE Third International Conference on Cloud and Green Computing. 2013. IEEE.

[14] Kern, R., H. Thies and G. Satzger, "Statistical quality control for human-based electronic services," Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2010, 6470 LNCS, (Mv), pp. 243-257.

[15] Raykar, V.C. and S. Yu, "Eliminating spammers and ranking annotators for crowdsourced labeling tasks," Journal of Machine Learning Research, 2012, 13, pp. 491-518.

[16] Alonso, O. and S. Mizzaro, "Using crowdsourcing for trec relevance assessment," Information Processing and Management, 2012, 48, (6), pp. 1053-1066.

[17] Samimi, P., S.D. Ravana, W. Webber and Y.S. Koh, "Effects of objective and subjective competence on the reliability of crowdsourced relevance judgments," Information Research, 2017, 22, (1).

[18] Alonso, O., "Implementing crowdsourcing-based relevance experimentation: An industrial perspective," Information Retrieval, 2013, 16, (2), pp. 101-120.

[19] Li, H. and Q. Liu, "Cheaper and better: Selecting good workers for crowdsourcing," in Third AAAI Conference on Human Computation and Crowdsourcing. 2015.

[20] Dang, D., Y. Liu, X. Zhang and S. Huang, "A crowdsourcing worker quality evaluation algorithm on mapreduce for big data applications," IEEE Transactions on Parallel and Distributed Systems, 2016, 27, (7), pp. 1879-1888.

[21] Welinder, P., S. Branson, S. Belongie and P. Perona, "The multidimensional wisdom of crowds," Advances in Neural Information Processing Systems 23: 24th Annual Conference on Neural Information Processing Systems 2010, NIPS 2010, 2010, pp. 1-9.

[22] Burnap, A., R. Gerth, R. Gonzalez and P.Y. Papalambros, "Identifying experts in the crowd for evaluation of engineering designs," Journal of Engineering Design, 2017, 28, (5), pp. 317-337.

[23] Li, K., S. Wang and X. Cheng, "Crowdsourcee evaluation based on persuasion game," Computer Networks, 2019, 159, pp. 1-9.

[24] Lucas, C., R.A. Nielsen, M.E. Roberts, B.M. Stewart, A. Storer and D. Tingley, "Computer-assisted text analysis for comparative politics," Political Analysis, 2015, 23, (2), pp. 254-277.

[25] Cheng, A.-S., K.R. Fleischmann, P. Wang, E. Ishita and D.W. Oard, "The role of innovation and wealth in the net neutrality—debate_ a content analysis of human values in congressional and fcc hearings," Journal of the American Society for Information Science and Technology, 2012, 64, (7), pp. 1360-1373.

[26] Quinn, K.M., B.L. Monroe, M. Colaresi, M.H. Crespin and D.R. Radev, "How to analyze political attention with minimal assumptions and costs," American Journal of Political Science, 2010, 54, (1), pp. 209-228.

[27] Mimno, D., H.M. Wallach, E. Talley, M. Leenders and A. McCallum, "Optimizing semantic coherence in topic models," EMNLP 2011 - Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference, 2011, (2), pp. 262-272.

[28] Graneheim, U.H. and B. Lundman, "Qualitative content analysis in nursing research: Concepts, procedures and measures to achieve trustworthiness," Nurse Education Today, 2004, 24, (2), pp. 105-112.

[29] Cao, B., X. Frank Liu, J. Liu and M. Tang, "Domain-aware mashup service clustering based on lda topic model from multiple data sources," Information and Software Technology, 2017, 90, pp. 40-54.

[30] Hagen, L., "Content analysis of e-petitions with topic modeling: How to train and evaluate lda models?," Information Processing and Management, 2018, 54, (6), pp. 1292-1307.

[31] Bastani, K., H. Namavari and J. Shaffer, "Latent dirichlet allocation (lda) for topic modeling of the cfpb consumer complaints," Expert Systems with Applications, 2019, 127, pp. 256-271.

[32] Jung, Y. and Y. Suh, "Mining the voice of employees: A text mining approach to identifying and analyzing job satisfaction factors from online employee reviews," Decision Support Systems, 2019, 123, (January), pp. 113074-113074.

[33] Ibrahim, N.F. and X. Wang, "A text analytics approach for online retailing service improvement: Evidence from twitter," Decision Support Systems, 2019, 121, (April), pp. 37-50.

[34] Mou, J., G. Ren, C. Qin and K. Kurcz, "Understanding the topics of export cross-border e-commerce consumers feedback: An lda approach," Electronic Commerce Research, 2019, 19, (4), pp. 749-777.

[35] Zhao, F., Y. Zhu, H. Jin and L.T. Yang, "A personalized hashtag recommendation approach using lda-based topic model in microblog environment," Future Generation Computer Systems, 2016, 65, pp. 196-206.

[36] Qin, G., D. Sanhong and W. Hao, "Chinese stopwords for text clustering: A comparative study," Data Analysis and Knowledge Discovery, 2017, 1, (03), pp. 72-80.
官琴, 邓三鸿, 王昊. 中文文本聚类常用停用词表对比研究[J]. 数据分析与知识发现. 2017,1(03):72-80.