Mathematical Problems of Computer Science 41, 55–62, 2014.

Engineering Competence Frameworks and Topic Modelling

Tigran R. Topchyan

Yerevan State University, Yerevan, Armenia
e-mail: tigran@topsoft.am

Abstract

Engineering education is often presented as frameworks made up of intended competences based on product lifecycle step models. Such are the Conceiving-Designing-Implementing-Operating(CDIO) syllabus and the European e-Competence Framework (e-CF). In this article we analyze the correlation between the expert created competences implemented in these models and the job market requirement qualifications represented as topics extracted with probabilistic topic models.

Keywords: LDA, Topics, Qualification, Competence, Framework.

1. Introduction

Universities are responsible for improving the quality of education and should address two central questions [1]:

- What is the full set of knowledge, skills, and attitudes that engineering students should possess as they leave the university, and at what level of proficiency?
- How can we do better at ensuring that students learn these skills?

Normally, at engineering universities, the internal quality assurance processes use the method of stakeholder focus groups comprised of engineering faculty, students, employers as industry representatives, university review committees, alumni, and senior academics. The focus group participants are asked questions designed to address these issues. This focus group and survey method is fundamental, but has a number of issues affecting the overall process scalability and outcome quality.

First, in context of globalization, organizing wide area surveys is very expensive and is impractical to carry out on a regular basis. Second, even if the necessary data is present, it can be difficult to recognize the dynamics of latent processes of changes in market requirements. Third, employers are not highly motivated to participate in this surveys as it demands effort and expenses on their part and there are no short-term outcomes. Fourth, the universities can be biased and so the overall data can be subjective.

In our previous work [2] we applied the probabilistic topic models to analyze a job description corpus, gathered from on-line job hiring sites, to recognize the qualification requirements sought by employers by extracting the underlying topics present in the data. These topics treated as qualifications proved to be useful and provided valuable insight into the job market. This method still requires an operator from the academia to manually sift through the data in order to benchmark the current qualification requirements of knowledge and skills and what amendments can be done to the current curriculum. This approach can be effective, because it is less expensive, can be used in a continuous manner in order to recognize the dynamics of market and it directly relates to the employers motivation to search for new talents.

Possible drawbacks of this method include the complexity of gauging the validity of the model and the presence of the human factor in the model analysis. To address both these problems we propose connecting our approach to the existing competences based models, such as the CDIO and e-CF frameworks [1, 3]. This way the validity of the model can be gauged in the context of how well it maps to the expert created models. Additionally, the labelling of the extracted topics with specific competences provides a more detailed yet simple approach to analyzing the model.

2. Topic Modelling for Competence Analysis

In our previous work we researched the issue of inferring qualifications and analyzing them from unstructured data. To this end, we experimented with probabilistic topic models for knowledge extraction, more specifically we used Latent Dirichlet Allocation (LDA) [4] to extract the underlying topical structure of a dataset of job descriptions. The trained model was able to effectively extract meaningful topics from the dataset, which could be mapped to requirements and qualifications that the job employers were looking for in potential employees. We outlined the power and flexibility of the extracted model and proposed a number of ways to analyze job description datasets, which could be useful in the context of qualification analysis.

In this work we aim to extend upon these results and research the connection between our extracted compact representations of a large amount of job descriptions and the established competency and qualification ontologies accepted as an industry standard. For this we again use LDA as our main method for topic extraction and below we present a short description of the key concepts we aim to exploit in our analysis.

3. LDA Primer

LDA maps collection of documents into a more compact topic based form. Topics in this case are probability distributions over a fixed vocabulary of terms. Each document in turn can be viewed as a vector of a subset of topics. These topics represent the thematic structure of all the documents.

LDA models the thematic structure using hidden random variables and tries to use these variables to create a model that best fits the collection of documents being analyzed. This can be viewed as a generative process of extracting topics which best match the documents in the dataset. This process is outlined below.

1. For each topic k in K sample a distribution β_k over the fixed vocabulary

2. For each document d in collection (D):

(a) Choose a distribution over K elements θ_d for the per-document topic proportions

(b) For each word w in d:

i. Sample a topic assignment z_n from θd , where z_n lies in $[1, K]$

ii. Choose a word w_n from z_n-th topic β_{z_n}

The key variables we can use for our analysis are the per-document topic proportions (θ_d), the word-topics proportions (β_k) and per-topic word assignments (z_n). In this particular case as we aim to explore the relation between the topics and existing competent ontologies, so we focus on β_k and z_n. Essentially they would supply us with a fixed number of topics describing the job descriptions and the proportions that each word appears in every topic.

The work in [5] outlines various methods that are used to infer the model based on this process, in this case we use [3] as we are building on system of continuous extraction of new job descriptions and it would allow for more convenient and fast updates to the model.

4. Competence Frameworks

In this work we are interested in the relation and mapping between our extracted qualifications sought by the employers and the more abstract representation of qualifications described in the existing competence ontologies. Here we review two of the more prominent engineering education competences- CDIO and e-CF [1, 6]. These frameworks use a life-cycle approach, which is the closest to employers' requirements, as the competence areas are latent by definition of the work organization.

CDIO. An approach that is currently dominant in the engineering field is CDIO, which is based on research of the industry and survey data and describes qualifications as competence areas, which contain competences describing the area. The CDIO syllabus in described in Table 1.

Table 1: CDIO condensed syllabus

1. Conceiving and Engineering	**2. Designing**
1. Setting System Goals and Requirements	1. The Design Process
2. Defining Function, Concept and Architecture	2. The Design Process Phasing and Approaches
3. Modelling of System and Ensuring Goals Can Be Met	3. Utilization of Knowledge in Design
4. Development Project Management	4. Disciplinary Design
	5. Multidisciplinary Design

3. Implementing	**4. Operating**
1. Designing the Implementation Process	1. Designing and Optimizing Operations
2. Hardware Manufacturing Process	2. Training and Operations
3. Software Implementing Process	3. Supporting the System Lifecycle
4. Hardware Software Integration	4. System Improvement and Evolution
5. Test, Verification, Validation, and Certification	5. Disposal and Life-End Issues
6. Implementation Management	6. Operations Management

e-CF. Another framework for qualification is European Competence Framework (e-CF), which focuses on the IT industry. e-CF is a well established framework that is based on research of the industry and models qualifications as five competence areas each containing descriptive competences:

1. **Plan**: contains nine competences
2. **Build**: contains six competences
3. **Run**: contains four competences
4. **Enable**: contains twelve competences
5. **Manage**: contains nine competences

5. From Topics to Competences

Using the techniques from our previous work [2] we can extract the topics that describe the thematic structure of the job descriptions. We proposed that these topics can be taken as representations of qualifications most sought after by employers. As already discussed, we want to find the connection between this extracted qualifications and the above-mentioned competence frameworks. This kind of analysis is useful as it can allow us to gauge the accuracy and coherence of the extracted topics in the context of expert created frameworks. Such a mapping can be used to analyze the framework and the job market in the context of the frameworks. Such a mapping can be used to find the most prominent competence currently sought after in the market and other uses, which becomes a grid for practical competences most focused by employers.

It is useful to find the connection between a topic and each of the separate competence areas described in the frameworks. This will provide an indirect labelling of the topics based on which area they are most probable to be included in. Each area of a framework is uniquely divided into competences, which are represented by concise textual descriptions (Table 1). These lend themselves well to be compared to our topics, which are distributions over words.

5.1 Mapping

As mentioned above we seek to find the probability of each topic being part of a specific competence. This problem cannot be efficiently solved by a supervised learning approach as it would require a one-to-many labelling by an expert. This would be impractical as we aim for a completely automated process. So to define the mapping we sought to find an empirical way to transform a topic into the context of a competence framework. For example, within the CDIO framework we seek to map the vector representing the topic into a compact vector of a distribution over the CDIO competence areas. Therefore, for topic β_k:
$\sum a_{a_b} + \beta$

$$\beta_k = [(w_1, p_1), ..., (w_l, p_l)] \longleftrightarrow C_k = [(c_{k_1}, p_1^*), ..., (c_{k_{areas}}, p_{areas}^*)] \text{ for } l \in [0, n], \qquad (1)$$

where C_k is the vector of pairs of the competences c proportions p^*.

To find C_k we need to find how similar each word in topics k is to each competence area. As each competence area is defined by separate competences, which are defined by words, we can treat the words describing the competences as text data we have to compare to each word w in topics β_k. Here we need to find how similar each word is to every competence area, by comparing how similar it is to sets of words describing every competence.

So, the for word w we will find the competence area c_w of that word by:

$$c_w = \max_{i \in (0, areas)} \left(\frac{\sum_{j=1}^{p} wnsim(w, c_{i,j})}{p} \right). \tag{2}$$

As in the previous work $wnsim$ is a similarity function defined over the WordNet graph, which compares the shortest distance between two words on the graph [7]. Using Eq.2 we can find the competence areas of all the words $(c_{k_{w_1}}, ..., c_{k_{w_l}})$ for the topic k and using that we can find the probability of competence area c_i using:

$$p^*_{c_i} = \left(\sum p_j | \forall j : c_{k_{w_j}} = c_i \right). \tag{3}$$

So we can find all the proportions $(p^*_1, ..., p^*_{areas})$. We can sum up the proportions of all the words mapped to the same competence as we would have arrived at the same LDA model if we have initially substituted that word by the competence area. This is due to LDA only considering documents word by word.

Therefore, we can arrive at a vector representation C_k from Eq.1, which will give us the representation of the topic on the space of competences. This method is the same for other competence frameworks and the analogous calculations for e-CF would hold.

6. Analysis

Using the method described in the previous section, we can project all the topics to competence space and explore even further. To this end, we trained an LDA model of 150 topics on an extended dataset of 10,000 job descriptions gathered in period from March-April 2014. The dataset contains preprocessed [2] job description data for the IT and Engineering industries. We then use the described method to project those topics into the CDIO and e-CF competence spaces. For each of the 150 topics we build a compact vector representation. We then train a simple Nearest Neighbour (NN) model [8] on that data in order to facilitate the robust analysis in-order to reveal relationships in the data. This would allow us to query for topics of arbitrary proportions. The results of the analysis are presented below.

CDIO. Using our NN model we found the closest topics for each of the competences in CDIO, which are depicted in Table 2. It can be seen that the extracted topics are indeed a certain representation of each competence. While not ideal, it proves that a topic model can be mapped to competences in an effective manner.

When analyzing the CDIO mapping we found that most topics are closer to the "Implementing" competence, as illustrated in Fig. 1. One can view this as a logical result, as it would reflect the general requirements for practical qualifications in the job market. But it may also be caused by a bias when scoring competences are based on WordNet distance, as this approach does not take context into account. This issue will be investigated in depth in our future work.

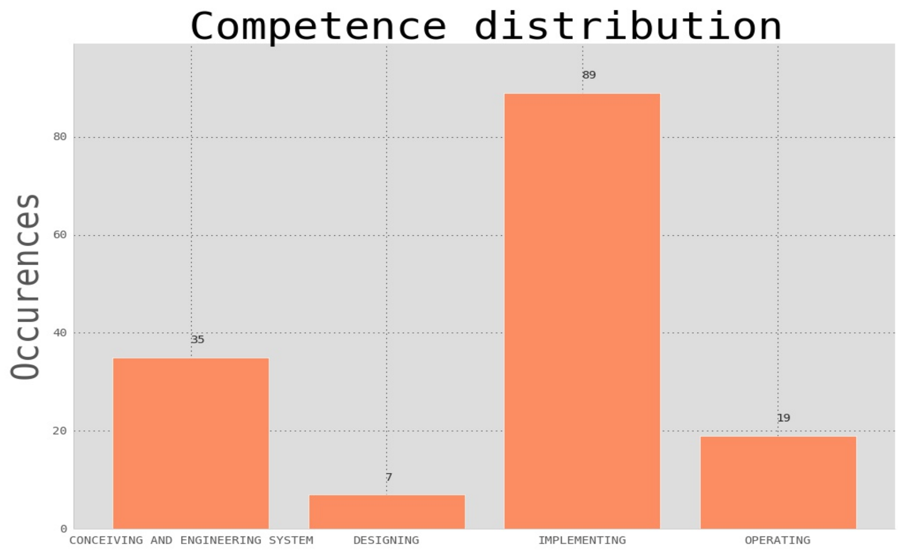

Fig. 1. The "Implementing" competence seems to be much more popular.

Table 2: Topics per competence for CDIO

1. Conceiving and Engineering	**2. Designing**
knowledge interpret procedures effectively skills basic responsibilities concepts environment personnel mathematical abilities analyze respond diagram policies worker qualifications travel techniques	manufacturing quality engineering tooling statistical skills procedures environment programs management manufacture technical reduction techniques developing achieve analyze tool knowledge leadership

3. Implementing	**4. Operating**
architect functional consultant analyst finance implementation environment reporting client phases technology planning global objectives guiding technical stability programs ensures objective	develops knowledge performs complex participates assignments applies implements maintains analyzes engineering functional tests conducts evaluates identifies technical coordinates demonstrates creates

Table 3: Topics per competence for e-CF.

1. Plan	**2. Build**
automation testing java automated test framework quality functional scripts plan	technology solutions technologies teams technical platforms engineering architects management infrastructure

3. Run	**4. Enable**
procedures management skills policies supervision knowledge personnel technical responsibilities engineering	technical management javascript java solutions customers developer agile visual developers

5. Manage
architect functional consultant align analyst finance implementation environment reporting client

E-CF. We carried out the same analysis for the e-CF competence framework and calculated the closest topics for each area (Table 3). It is evident that the topics closest to the competences are not very descriptive for this case. We attribute this to the descriptions of each area in e-CF being general and vague if taken as individual words. As mentioned above we are working on finding metrics for similarity that take context into account.

7. Summary

We presented a method for analyzing the qualifications extracted from job description data in the context of existing competence frameworks. The encouraging preliminary results showed that the topics extracted using the topic modelling can, indeed, be mapped to expert created concepts for qualifications and competences. It provides a novel approach for researchers to analyze the job market. This can be an invaluable tool for competence and qualification research in the future.

References

[1] E. F. Crawley, J. Malmqvist, W. A. Lucas and D. R. Brodeur, "The cdio syllabus v2. 0. an updated statement of goals for engineering education", In *Proceedings of 7th International CDIO Conference*, Copenhagen, Denmark, 2011.

[2] T. Topchyan, "Job Market requirements and qualities extraction for qualification analysis", *Transactions of IIAP NAS RA, Mathematical Problems of Computer Science*, vol. 41, pp. 5–14, 2014.

[3] M. D. Hoffman, D. M. Blei and F. R. Bach, "Online learning for latent dirichlet allocation", *Neural Information Processing Systems (NIPS)*, Curran Associates, Inc., pp. 856–864, 2010.

[4] D. M. Blei, A. Y. Ng and M. I. Jordan, "Latent dirichlet allocation", *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, March 2003.

[5] D. M. Blei. J.D. Lafferty, "Topic Models", *Text Mining: Classification, Clustering, and Applications, ed. Ashok Srivastava and Mehran Sahami. London, England: Taylor and Francis*, pp. 71–94, 2009.

[6] "European e-Competence Framework 3.0", 2014.

[7] A. Budanitsky and G. Hirst, "Evaluating wordnet-based measures of lexical semantic relatedness", *Computational Linguistics*, vol. 32, no. 1, pp. 13–47, March 2006.

[8] S. Arya and D. M. Mount, "Algorithms for fast vector quantization", *Proceedings of the IEEE Data Compression Conference*, pp. 381–390, 1993.

Submitted 26.12.2013, accepted 28.02.2014.

Ճարտարագիտական կրթության չափորոշիչներ և թեմաների մոդելավորում

Տ. Թոփչյան

Ամփոփում

Ճարտարագիտական կրթության չափորոշիչները հաճախ կառուցվում են արտադրական գործընթացի ամբողջական շրջափուլի հիմքի վրա: Այդպիսի չափորոշիչներ են հանդիսանում Մտահղացում-Մշակում-Իրականացում-դործարկում (CDIO) կրթական ծրագրի մոդելը և ՏՏ-Կոմպետենցիաների Եվրոպական Շրջանակը (e-CF): Ներկայացվող աշխատանքում մենք վերլուծում ենք փորձագետի կողմից մշակված չափորոշիչներում ներկայացված կոմպետենցիաների և հավանականային թեմաների մոդելի օգնությամբ դուրս բերված որակավորումների միջև եղած կապը:

Инженерные образовательные стандарты и тематическое моделирование

Т. Топчян

Аннотация

Инженерные образовательные стандарты часто строятся на основе комптенций необходимых на этапах полного жизненного цикла продукта производства. Таковыми являются Задумка-Проект-Реализация-Эксплуатация (CDIO) и Европейская Рамка ИКТ-Компетенций (e-CF). В данной статье предлагается анализ связи между создаваемыми экспертами рамок компетенций и квалификационными требованиями, выделенными с помощью вероятностной модели тем.