MARKET-REQUIRED COMPETENCE TOPIC DYNAMICS

Tigran Topchyan

Abstract: *The Job market is an ever moving and evolving entity. So are the competencies and qualifications it demands of prospective employees. In our previous work we modelled these competencies using topic models, but in order to have a more effective understanding of the market, we have to take into account the dynamics of the system as well. Here we propose using dynamic topic modelling as a means of analysing the job market for competencies and qualifications stemming from our previous research.*

Keywords: *dynamic topic modelling, qualification, competence, dynamics*

ACM Classification Keywords: *H.0 Information Systems-General*

Introduction

In our previous work we researched the issue of inferring qualifications and analysing them from unstructured data. To this end, we experimented with probabilistic topic models for knowledge extraction [Topchyan, 2014] 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 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 analyse job description datasets, which could be useful in the context of qualification analysis.

We also extended upon these results and researched the connection between our extracted representations and established competency and qualification ontologies accepted as an industry standard [Topchyan, 2014].

While such an analysis is useful, the job market and as a result competency and qualification requirements, are in constant motion and evolve over time. Some skills rise in importance, some decline. This constant shift is very important to build a comprehensive understanding of the industry requirements. To build an informative model of the competencies and qualifications we would require up-to-date information. For this job description are again an ideal means of analysis, as they are updated daily and shadow the dynamics and evolution of the competencies and qualification requirements. And if we consider the stream of job descriptions as time slices of document corpora we can again apply topic modelling to not only find the topics of the document corpora, but the topic dynamics as well, the topic change over time.

Dynamic Topic Models

In our previous work we only considered static topic models, which models a particular snapshot of a document collection without taking time into account. On the other hand, dynamic topic models view the document collection as slices of documents separated based on time-stamp. Several dynamic topic models have been proposed in the literature [Wang, 2006],[Zhang, 2006]. Here we propose one of the first approaches due to its similarity to the method used in our previous work [Topchyan, 2014] and its proven performance. We could also use standard static LDA [Blei, 2003], but retrained for each slice, but [Blei, 2006] shows that this approach is more effective.

The dynamic topic model (DTM) , as described in [Blei, 2006] models the document corpus as being separated into time slices, which contain a certain set amount of documents. The particular time slices can be any equally spaced out time denominations - hours, days, months, years. Contrary to static LDA [Blei, 2003], where the K topics were

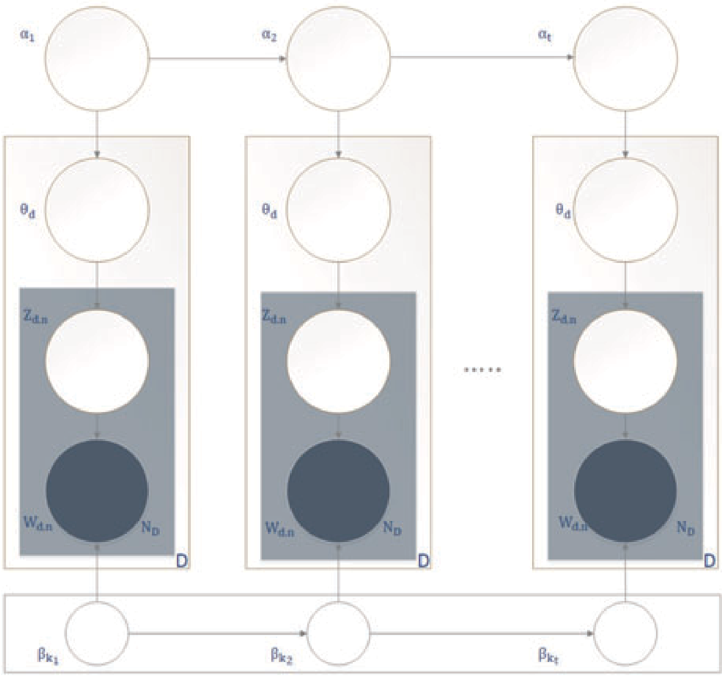

Figure 1: Graphical model for DTM.

described as Dirichlet distribution over a fixed vocabulary V of words β_k, DTM chains the natural parameters of each topic $\beta_{t,k}$ in a state space model that evolves with Gaussian noise;

$$\beta_{t,k}|\beta_{t-1,k} \sim \mathcal{N}(\beta_{t-1,k}, \sigma^2 I) \tag{1}$$

The document-specific topic proportions θ are modelled in a similar fashion by using a logistic normal with mean α to express uncertainty over proportions;

$$\alpha_{t,k}|\alpha_{t-1,k} \sim \mathcal{N}(\alpha_{t-1,k},\ \delta^2 I) \tag{2}$$

The rest of the model can be interpreted as local instances of LDA with the parameters α and β propagating trough time. The generative process itself for a document collection of D document and K topics with t slices. It can be summarised as follows:

1. For topic k in K draw $\beta_{t,k}|\beta_{t-1,k} \sim \mathcal{N}(\beta_{t-1,k}, \sigma^2 I)$
2. Draw $\alpha_{t,k}|\alpha_{t-1,k} \sim \mathcal{N}(\alpha_{t-1,k},\ \delta^2 I)$
3. For document d in D
   (a) Draw $\eta \sim \mathcal{N}(\alpha_t, \alpha^2 I)$
   (b) For each word:
      i. $Z \sim Mult(\pi(\eta))$
      ii. $W_{t,d,n} \sim Mult(\pi(\beta_{t,z}))$

The generative process is also summarised as a graphical model in Figure. 1.

The model inference is an extension of the ideas from the static models. In this case variational Kalman filtering was used. The time slices were all initially initialized as local LDA models and then the filtering was applied for accurate posterior inference [Blei, 2006] .

Analysis

We applied the Dynamic Topic Model on our corpus of ten thousand gathered documents from the IT industry. The documents were gathered across a span six weeks. We trained two models one with 20 topics and one with 100 topics. The training time took 5 and 15 hours respectively. In this section we present the results for the 20 topic model.

Our system supports gathering document from a multitude of sources, but for this experiment we limited it to one in order to gauge the performance and viability of the model for our purposes. Generally the purposed timespan will not contain drastic change, but nonetheless we want to present our findings on the applicability of such a model in the context of competence and qualification analysis.

The DTM is a very robust model and gives us not only the topics in time, but also allows us to infer the document influence model. Meaning we can find how influential each document was to the change in the topics across time. This can potentially have very interesting applications in our research, in particular in the discovery of turning points of competence. But in this work we are more interested in the per-topic dynamics.

Per-Topic: In the context of dynamics analysis we are particularly interested in how each topic changes in time. We are interested in how the topics change in time, meaning how the probabilities of each word change with each time-step. We analysed our corpus and using concepts presented in our previous work [Topchyan, 2014] found the more coherent and popular topics in the dataset and analysed them.

We present the dynamics as a plot of the change of probabilities of four of the most salient words of the topic as well as an embedded wordcloud timeline representation. The size of the words in the wordcloud corresponds to their probability in the given topic. An example of a topic related to practical competences related to databases and problem solving skills can be seen in Figure 2.

From Figure 2 we can see the dynamics of the topic. As we mentioned the time-slices are too close together for any drastic changes to occur, but nonetheless we can notice the steady increase in probability for most of the terms, signifying the fact that the separate time slices contained job descriptions with requirements along the lines of the ones already seen previously. This is fairly logical as the these competencies are fairly popular and sought after in the modern industry and are fairly practical so they occur more often.

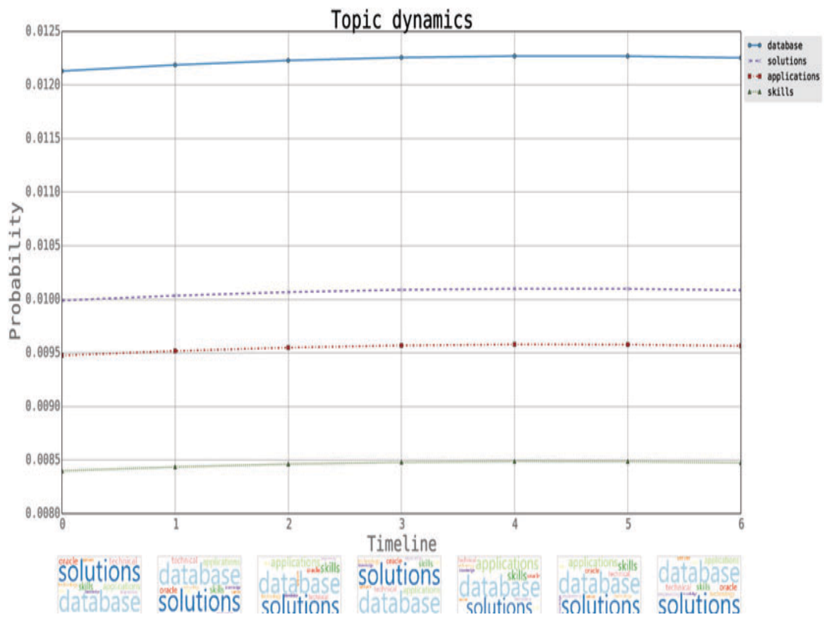

Figure 2: Topic dynamics for database and problem solving competences.

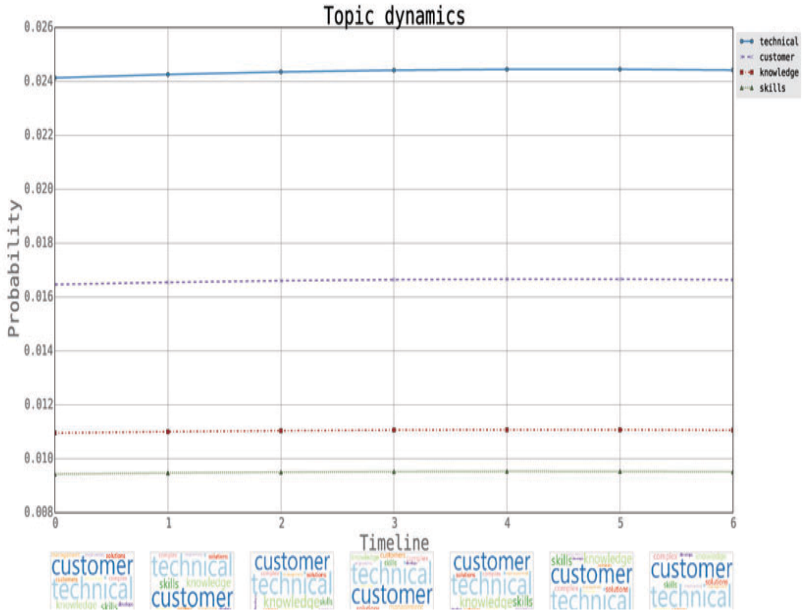

Figure 3: Topic dynamics peronsal and customer relation competences.

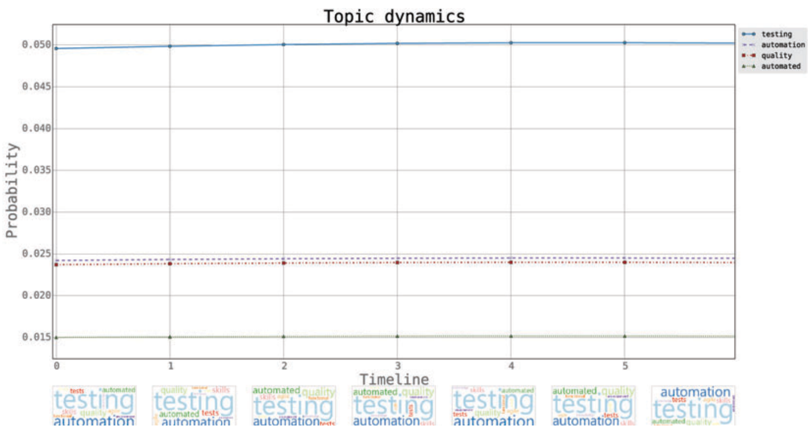

Figure 4: Topic dynamics for topic, that fits perfectly in the CDIO 'IMPLEMENTING' competency group [Crawley, 20], as discussed in our previous work .

On the other hand the dynamics in Figure 3 show that the more probable term start to decrease in probability after a number of time-steps. This can interpreted as the topic contains competences related to more abstract knowledge related to working with customers, and it stands to reason, that in the very practical IT industry there might be a certain drop off in such competence requirements.

Figure 4 present a very interesting result concerning a topic, that is directly related to a CDIO competency group [Crawley, 20], specifically the "Test, Verification, Validation and Certification" competency. We can see that the probabilities rise and do not change thought the timestep. We explain this by the fact, that this is a very sought after competency in the industry and virtually every job description usually contains something concerning testing and verification. This is an encouraging result and proves, that the model is useful for inferring insight from the data.

Conclusion

We researched the viability of using dynamic topic models for analysing competencies and qualification requirements from job market data. The encouraging preliminary results showed that topics extracted using topic modelling can be interpreted to reflect the dynamics of the market and the sought after competencies. In conjunction with out previous work on competency topic modelling this can lead to invaluable contributions to competency and qualification research in the future.

Bibliography

[Blei, 2003] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. In: Latent dirichlet allocation. J. Mach. Learn. Res, 3:993–1022, mar 2003.

[Blei, 2006] Blei, David M and Lafferty, John D. In: Dynamic topic models. Proceedings of the 23rd international conference on Machine learning, 113–12, 2006.

[Wang, 2006] Wang, Chong and Blei, David and Heckerman, David. In: Continuous time dynamic topic models. arXiv preprint arXiv:1206.3298, 2012.

[Zhang, 2006] Zhang, Jianwen and Song, Yangqiu and Zhang, Changshui and Liu, Shixia. In: Evolutionary hierarchical dirichlet processes for multiple correlated time-varying corpora. Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, 1079–1088, 2010.

[Topchyan, 2014] Tigran Topchyan. In: Engineering competence frameworks and topic modelling. Mathematical Problems of Computer Science, 2014.

[Topchyan, 2014] Tigran Topchyan. In: Job Market requirements and qualities extraction for qualification analysis. Mathematical Problems of Computer Science, 2014.

[Crawley, 20] Edward F Crawley, Johan Malmqvist, William A Lucas, and Doris R Brodeur. In: The cdio syllabus v2. 0 an updated statement of goals for engineering education. Proceedings of 7th International CDIO Conference, Copenhagen, Denmark, 2011.

Authors' Information

Tigran Topchyan *Yerevan State University, Yerevan, Armenia; e-mail: tigran@topsoft.am*
Major Fields of Scientific Research: Machine Learning
Education organization