A Probabilistic Model for Knowledge Component Naming

Cyril Goutte
National Research Council
1200 Montreal Rd
Ottawa, ON, Canada
Cyril.Goutte@gmail.com

Serge Léger
National Research Council
100 rue des Aboiteaux
Moncton, NB, Canada
Serge.Leger@nrc.ca

Guillaume Durand
National Research Council
100 rue des Aboiteaux
Moncton, NB, Canada
Guillaume.Durand@nrc.ca

ABSTRACT

Recent years have seen significant advances in automatic identification of the Q-matrix necessary for cognitive diagnostic assessment. As data-driven approaches are introduced to identify latent knowledge components (KC) based on observed student performance, it becomes crucial to describe and interpret these latent KCs. We address the problem of naming knowledge components using keyword automatically extracted from item text. Our approach identifies the most discriminative keywords based on a simple probabilistic model. We show this is effective on a dataset from the PSLC datashop, outperforming baselines and retrieving unknown skill labels in nearly 50% of cases.

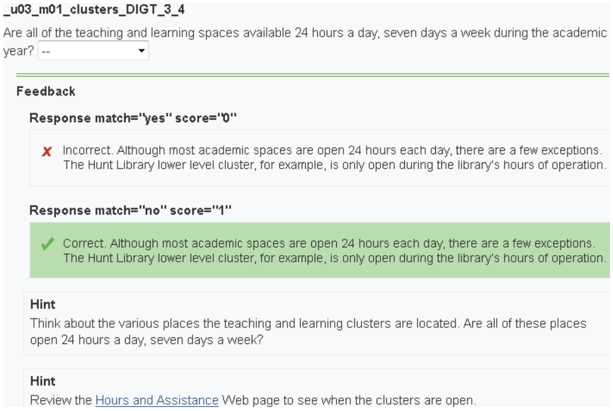
1. OVERVIEW

The Q-matrix, introduced by Tatsuoka [9], associates test items with attributes of students that the test intends to assess. A number of data-driven approaches were introduced to automatically identify the Q-matrix by mapping items to latent *knowledge components* (KCs), based on observed student performance [1, 6], using, e.g. matrix factorization [2, 8], clustering [5] or sparse factor analysis [4]. A crucial issue with automatic methods is that latent skills may be hard to describe and interpret. Manually-designed Q-matrices may also be insufficiently described. A data-generated description is useful in both cases.

We propose to extract *keywords* relevant to each KC from the textual content corresponding to each item. We build a simple probabilistic model, with which we score keywords. This proves surprisingly effective on a small dataset obtained from the PSLC datashop.

2. MODEL

We focus on extracting keywords from the textual content of each item (question, hints, feedback, Fig. 1). We denote by d_i the textual content (e.g. body text) of item i, and assume a Q-matrix mapping items to K skills c_k, $k = 1 \dots K$.



_u03_m01_clusters_DIGT_3_4

Are all of the teaching and learning spaces available 24 hours a day, seven days a week during the academic year? --

Feedback

Response match="yes" score="0"

✗ Incorrect. Although most academic spaces are open 24 hours each day, there are a few exceptions. The Hunt Library lower level cluster, for example, is only open during the library's hours of operation.

Response match="no" score="1"

✓ Correct. Although most academic spaces are open 24 hours each day, there are a few exceptions. The Hunt Library lower level cluster, for example, is only open during the library's hours of operation.

Hint

Think about the various places the teaching and learning clusters are located. Are all of these places open 24 hours a day, seven days a week?

Hint

Review the Hours and Assistance Web page to see when the clusters are open.

Figure 1: Example item body, feedback and hints.

These may be latent skills obtained automatically or from a manually designed Q-matrix. For each KC we build a unigram language model estimating the relative frequency of words in each KC [7]:

$$P(w|c_k) \propto \sum_{i, d_i \in c_k} n_{wi}, \qquad \forall k \in \{1 \dots K\} \tag{1}$$

with n_{wi} the number of occurrences of word w in document d_i. $P(w|c)$ is the *profile* of c. Important words are those that are high in c's profile and low in other profiles. The symmetrized Kullback-Leibler divergence between $P(w|c)$ and the profile of all other classes, $P(w|\neg c)$, decomposes over words: $KL(c, \neg c) = \sum_w \left(P(w|c) - P(w|\neg c)\right) \log \frac{P(w|c)}{P(w|\neg c)}$. We use the contribution of each word to the KL divergence as score indicative of keywords. In order to focus on words significantly *more* frequent in c, we use the signed score:

$$\text{KL score:} \quad s_c(w) = |P(w|c) - P(w|\neg c)| \log \frac{P(w|c)}{P(w|\neg c)}. \tag{2}$$

Figure 2 illustrates this graphically. Words frequent in c but not outside (green, right) receive high positive scores. Words rare in c but frequent outside (red, left) receive negative scores. Words equally frequent in c and outside (blue) get scores close to zero: they are not specific enough.

3. EXPERIMENTAL RESULTS

We used the 100 student random sample of the ”Computing@Carnegie Mellon” dataset, *OLI C@CM v2.5 - Fall 2013, Mini 1*. This OLI dataset is well suited for our study because the full text of the items is available in HTML format

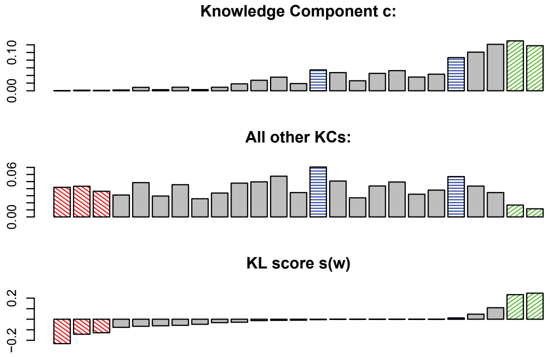


Figure 2: From KC profile, other KCs, to KL scores.

KC label	#it	Top 10 keywords (body text only)
identify-sr	52	phishing email scam social learned indicate legitimate engineering anti-phishing indicators
print quota	12	quota printing andrew print semester consumed printouts longer unused cost
penalties bandwidth	1	maximum limitations exceed times bandwidth suspended network access

Table 1: Top 10 keywords for 3 KC of various sizes.

and can be extracted. Other datasets only include screenshots. There are 912 unique steps, 31k body tokens, 11.5k hints tokens, and 41k feedback tokens, close to 84k tokens total. We pick a model in PSLC that has 108 distinct KCs with partially descriptive labels. That model assigns 1 to 52 items to each KC, for 823 items with at least 1 KC assigned. All text is tokenized, stopwords are removed, as well as tokens not containing one alphabetical character.

We estimate three different models, using Eq. (1), depending on the data considered: body text only ("body"), body and hints ("b+h"), all text ("all"). For each model, we extract up to 10 words with highest KL score (2) for each KC. Table 1 shows that even for knowledge components with very few items, the extracted keywords are clearly related to the topic suggested by the label. Although the label itself is not available when estimating the model, words from the label often appear in the keywords: this happens in 44 KCs out of 108 (41%), suggesting that the retrieved keywords are relevant. Note that some labels are vague (e.g. *identify-sr*) but the keywords provide a clear description (*phishing scams*).

We now focus on two desirable qualities for good keywords: *diversity* (keywords should differ accross KCs) and *specificity* (keywords should describe few KCs). Table 2 compares KL scores with the common strategy of picking the most frequent words (MP), using various metrics. Good descriptions should have a high number of different keywords, many of which describing a unique KC, and few KCs per keyword. The total number of keyword is fairly stable as we extract up to 10 keywords for 108 KCs. It is clear that KL extracts many more different keywords (up to 727) than MP (352 to 534). KL yields on average 1.4 (median 1) KC per keyword, whereas MP keywords describe on average 3.1 KC. There are also many more KL-generated keywords describing a unique KC. These results support the conclusion that our KL-based method provides better *diversity* and *specificity*.

	total	different	unique	max
KL-body	995	**727**	**577**	**9**
KL-b+h	1005	722	558	10
KL-all	1080	639	480	19
MP-body	995	534	365	42
MP-b+h	1005	521	340	34
MP-all	1080	352	221	87

Table 2: Keyword extraction for KL vs. max. probability (MP) using text from body, b+h and all fields; total keywords, # different keywords, # with unique KC, and maximum KC per keyword.

Note that using more textual content (adding hints and feedback) hurts performance accross the board. We see why from the list of words describing most KCs from two methods:
KL-body: use (9) following (8) access, andrew, account (7)
MP-all: incorrect(87) correct(67) review(49) information(30)

"correct" and "incorrect" are extracted for 67 and 87 KCs, respectively, because they appear frequently in the feedback text. The KL-based approach discards them because they are equally frequent everywhere.

Acknowledgement

We used the 'OLI C@CM v2.5 - Fall 2013, Mini 1 (100 students)' dataset accessed via DataShop [3]. We thank Alida Skogsholm from CMU for her help in choosing this dataset.

4. REFERENCES

[1] T. Barnes. The Q-matrix method: Mining student response data for knowledge. In *AAAI EDM workshop*, 2005.

[2] M. Desmarais. Mapping questions items to skills with non-negative matrix factorization. *ACM-KDD-Explorations*, 13(2), 2011.

[3] K.R. Koedinger, R.S.J.d. Baker, K. Cunningham, A. Skogsholm, B. Leber, and J. Stamper. A data repository for the EDM community: The PSLC datashop. In *Handbook of Educational Data Mining.* CRC Press, 2010.

[4] A.S. Lan, C. Studer, and R.G. Baraniuk. Quantized matrix completion for personalized learning. In *7th EDM*, 2014.

[5] N. Li, W. Cohen, and K.R. Koedinger. Discovering student models with a clustering algorithm using problem content. In *6th EDM*, 2014.

[6] J. Liu, G. Xu, and Z. Ying. Data-driven learning of Q-matrix. *Applied Psych. Measurement*, 36(7), 2012.

[7] A. McCallum and K. Nigam. A comparison of event models for naive bayes text classification. In *AAAI-98 workshop on learning for text categorization*, 1998.

[8] Y. Sun, S. Ye, S. Inoue, and Yi Sun. Alternating recursive method for q-matrix learning. In *7th EDM*, 2014.

[9] K.K. Tatsuoka. Rule space: an approach for dealing with misconceptions based on item response theory. *J. of Educational Measurement*, 20(4), 1983.