Measuring Similarity of Educational Items Using Data on Learners' Performance

Jiří Řihák
Faculty of Informatics
Masaryk University
Brno, Czech Republic
thran@mail.muni.cz

Radek Pelánek
Faculty of Informatics
Masaryk University
Brno, Czech Republic
pelanek@mail.muni.cz

ABSTRACT

Educational systems typically contain a large pool of items (questions, problems). Using data mining techniques we can group these items into knowledge components, detect duplicated items and outliers, and identify missing items. To these ends, it is useful to analyze item similarities, which can be used as input to clustering or visualization techniques. We describe and evaluate different measures of item similarity that are based only on learners' performance data, which makes them widely applicable. We provide evaluation using both simulated data and real data from several educational systems. The results show that Pearson correlation is a suitable similarity measure and that response times are useful for improving stability of similarity measures when the scope of available data is small.

1. INTRODUCTION

Interactive educational systems offer learners items (problems, questions) for solving. Realistic educational systems typically contain a large number of such items. This is particularly true for adaptive systems, which try to present suitable items for different kinds of learners. The management of a large pool of items is difficult. However, educational systems collect data about learners' performance and the data can be used to get insight into item properties. In this work we focus on methods for computing item similarities based on learners' performance data, which consists of binary information about the answers (correct/incorrect).

Automatically detected item similarities are the first and necessary step in further analysis such as clustering of the items, which is useful in several ways, with one particular application being learner modeling [9]. Learner models estimate knowledge and skills of learners and are the basis of adaptive behavior of educational systems. A learner's models requires a mapping of items into knowledge components [17]. Item clusters can serve as a basis for knowledge component definition or refinement. The specified knowledge components are relevant not only for modeling, but they are typically directly visible to learners in the user interface of a system, e.g., in a form of open learner model visualizing the estimated knowledge state, or in a personalized overview of mistakes, which is grouped by knowledge components.

Information about items is also very useful for management of the content of educational systems – preparation of new items, filtering of unsuitable items, preparation of explanations, and hint messages. Information about item similarities and clusters can be also relevant for teachers as it can provide them an inspiration for "live" discussions in class. This type of applications is in line with Baker's argument [1] for focusing on the use of learning analytics for "leveraging human intelligence" instead of its use for automatic intelligent methods.

Item similarities and clusters are studied not only in educational data mining but also in a closely related area of recommender systems. The setting of recommender systems is in many aspects very similar to educational systems – in both cases we have users and items, just instead of "performance" (the correctness of answers, the speed of answers) recommender systems consider "ratings" (how much a user likes an item). Item similarities and clustering techniques have thus been also considered in the recommender systems research (we mention specific techniques below). There is a slight, but important difference between the two areas. In recommender systems item similarities and clusterings are typically only auxiliary techniques hidden within a "recommendation black box". In educational system, it is useful to make these results explicitly available to system developers, curriculum production teams, or teachers.

There are two basic approaches to dealing with item similarities and knowledge components: a "model based approach" and an "item similarity approach". The basic idea of the model based approach is to construct a simplified model that explains the observed data. Based on a matrix of learners' answers to items we construct a model that predicts these answers. Typically, the model assigns several latent skills to learners and uses a mapping of items to corresponding latent factors. This kind of models can often be naturally expressed using matrix multiplication, i.e., fitting a model leads to matrix factorization. Once we fit the model to data, items that have the same value of a latent factor can be denoted as "similar". This approach leads naturally to multiple knowledge components per skill. The model is typically computed using some optimization technique that leads only to local optima (e.g., gradient descent). It is thus necessary to address the role of initialization, and parameter setting of the search procedure. In recommender systems this approach is used for implementation of collaborative filtering; it is often called "singular value decomposition" (SVD) [18]. In educational context many variants of this approach have been proposed under different names and terminology, e.g., Q-matrix [3], non-negative matrix factorization techniques [8], sparse factor analysis [19], or matrix refinement [10].

With the item similarity approach we do not construct an explicit model of learners' behavior, but we compute directly a similarity measure for each pairs of items. These similarities are then used to compute clusters of items, to project items into a plane, or for other analysis (e.g., for each item listing the 3 most similar items). This approach naturally leads to a mapping with a single knowledge component per item (i.e., different kind of output from most model based methods). One advantage of this approach is easier interpretability. In recommender system research this approach is called neighborhood-based methods [11] or item-item collaborative filtering [7]. Similarity has been used for clustering of items [23, 24] and also for clustering of users [29]. In educational setting item similarity has been analyzed using correlation of learners' answers [22] and problem solving times [21], and also using learners' wrong answers [25].

So far we have discussed methods that are based only on data about learners' answers. Often we have some additional information about items and their similarities, e.g., a manual labeling or data based on syntactic similarity of items (text of questions). For both model based and item similarity approaches previous research has studied techniques for combination of these different types of inputs [10, 21].

In this work we focus on the item similarity approach, because in the educational setting this approach is less explored than the model based approach. We discuss specific techniques, clarify details of their usage, and provide evaluation using both data from real learners and simulated data. Simulated data are useful for evaluation of the considered unsupervised machine learning tasks, because in the case of real-world data we do not know the "ground truth".

The specific contributions of this work are the following. We provide guidelines for the choice of item similarity measures – we discuss different options and provide results identifying suitable measures (Pearson, Yule, Cohen); we also demonstrate the usefulness of "two step similarity measures". We explore benefits of the use of response time information as supplement to usual information of correctness of answer. We use and discuss several evaluation methods for the considered tasks. We specifically consider the issue of "how much data do we need". This is often practically more important than the exact choice of a used technique, but the issue is rather neglected in previous work.

2. MEASURES OF ITEM SIMILARITY

Figure 1 provides a high-level illustration of the item similarity approach. This approach consist of two steps that are to a large degree independent. At first, we compute an item similarity matrix, i.e., for each pair of items i, j we compute similarity s_{ij} of these items. At second, we can construct clusters or visualizations of items using only the item similarity matrix.

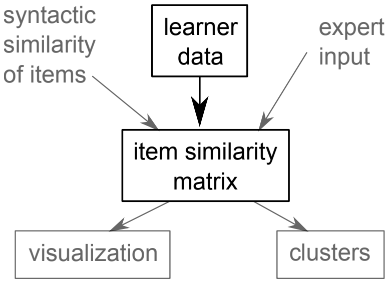

Figure 1: High-level illustration of the general approach to item analysis based on item similarities.

Experience with clustering algorithms suggests that the appropriate choice of similarity measure is more important than choice of clustering algorithm [13]. The choice of similarity measure is domain specific and it is typically not explored in general research on clustering. Therefore, we focus on the first step – the choice of similarity measure – and explore it for the case of educational data.

2.1 Basic Setting

In this work we focus on computing item similarities using learners' performance data. As Figure 1 shows, the similarity computation can also utilize information from domain experts or automatically determined information based on the inner structure of items (e.g., text of questions or some available meta-data).

We discuss different possibilities for computation of item similarities. Note that in our discussion we consistently use "similarity measures" (higher values correspond to higher similarity), some related works provide formulas for dissimilarity measures (distance of items; lower values correspond to higher similarity). This is just a technical issue, as we can easily transform similarity into dissimilarity by subtraction.

The input to item similarity computation are data about learner performance, i.e., a matrix $L \times I$, where L is the number of learners and I is the number of items. The matrix values specify learners' performance. The matrix is typically very sparse (many missing values). The output of the computation is an item similarity matrix, which specifies similarity for each pair of items.

Note that in our discussion we mostly ignore the issue of learning (change of learners skill as they progress through items). When learning is relatively slow and items are presented in a randomized order, learning is just a reasonably small source of noise and does not have a fundamental impact on the computation of item similarities. In cases where learning is fast or items are presented in a fixed order, it may be necessary to take learning explicitly into account.

2.2 Correctness of Answers

The basic type of information available in educational systems is the correctness of learners' answers. So we start with similarity measures that utilize only this type of information, i.e., dichotomous data (correct/incorrect) on learners' answers on items. The advantage of these measures is that they are applicable in wide variety of settings.

With dichotomous data we can summarize learners' performance on items i and j using an agreement matrix with just four values (Table 1). Although we have just four values to quantify the similarity of items i and j, previous research has identified large number of different measures for dichotomous data and analyzed their relations [5, 12, 20]. For example Choi et al. [5] discuss 76 different measures, albeit many of them are only slight variations on one theme. Similarity measures over dichotomous data are often used in biology (co-occurrence of species) [14]. A more directly relevant application is the use of similarity measures for recommendations [30]. Recommender systems typically use either Pearson correlation or cosine similarity for computation of item similarities [11], but they consider richer than binary data.

Table 1: An agreement matrix for two items and definitions of similarity measures based on the agreement matrix ($n = a + b + c + d$ is the total number of observations).

		item i	
		incorrect	correct
item j	incorrect	a	b
	correct	c	d

Yule	$S_y = (ad - bc)/(ad + bc)$
Pearson	$S_p = (ad - bc)/\sqrt{(a+b)(a+c)(b+d)(c+d)}$
Cohen	$S_c = (P_o - P_e)/(1 - P_e)$ $P_o = (a + d)/n$ $P_e = ((a + b)(a + c) + (b + d)(c + d))/n^2$
Sokal	$S_s = (a + d)/(a + b + c + d)$
Jaccard	$S_j = a/(a + b + c)$
Ochiai	$S_o = a/\sqrt{(a+b)(a+c)}$

Table 1 provides definitions of 6 measures that we have chosen for our comparison. In accordance with previous research (e.g., [5, 14]) we call measures by names of researchers who proposed them. The choice of measures was done in such a way as to cover measures used in the most closely related work and measures which achieved good results (even if the previous work was in other domains). We also tried to cover different types of measures.

Pearson measure is the standard Pearson correlation coefficient evaluated over the dichotomous data. In the context of dichotomous data it is also called Phi coefficient or Matthews correlation coefficient. *Yule* measure is similar measure, which achieved good results in previous work [30]. *Cohen* measure is typically used as a measure of inter-rater agreement (it is more commonly called "Cohen's kappa"). In our setting it makes sense to consider this measure when we view learners' answers as "ratings" of items. Relations between these three measures are discussed in [32].

Ochiai coefficient is typically used in biology [14]. It is also equivalent to cosine similarity evaluated over dichotomous data; cosine similarity is often used in recommender systems for computing item similarity, albeit typically over interval data [7]. *Sokal* measure is also called Sokal-Michener or "simple matching". It is equivalent to accuracy measure used in information retrieval. Together with *Jaccard* measure they are often used in biology, but they have also been used for clustering of educational data [12].

Note that some similarity measures are asymmetric with respect to 0 and 1 values. These measures are typically used in contexts where the interpretation of binary values is presence/absence of a specific feature (or observation). In the educational context it is more natural to use measures which treat correct and incorrect answers symmetrically. Nevertheless, for completeness we have included also some of the commonly used asymmetric measures (Ochiai and Jaccard). In these cases we focus on incorrect answers (value a as opposed to d) as these are typically less frequent and thus bear more information.

2.3 Other Data Sources

The correctness of answers is the basic source of information about item similarities, but not the only one. We can also use other data. The second major type of performance data is response time (time taken to answer an item). The basic approach to utilization of response time is to combine it with the correctness of an answer. Given the correctness value $c \in \{0, 1\}$, a response time $t \in \mathbb{R}^+$, and the median of all response times τ, we combine them into a single score r. Examples of such transformations are: linear transformation for correct answers only ($r = c \cdot max(1 - t/2\tau,\ 0)$); exponential discounting used in MatMat [28] ($r = c \cdot min(1,\ 0.9^{t/\tau - 1})$); linear transformation inspired by *high speed, high stakes scoring rule* used in Math Garden [16] ($r = (2c - 1) \cdot max(1 - t/2\tau,\ 0)$). The first approach was used in our experiment due to its simplicity and high influence of response time information.

The scores obtained in this way are real numbers. Given the scores it is natural to compute similarity of two items using Pearson correlation coefficient of scores (over learners who answered both items). It is also possible to utilize specific wrong answers for computation of item similarity [25].

It is also possible to combine performance based measures with other types of data. For example we may estimate item similarity based on analysis of the content of items (syntactical similarity of texts), or collect expert opinion (manual categorization of items into several groups). The advantage of the similarity approach (compared to model based approach) is that different similarity measures can be usually combined in straightforward way by using a weighted average of different measures.

2.4 Second Level of Item Similarity

The basic computation of item similarities computes similarity of items i and j using only data about these two items. To improve a similarity measure, it is possible to employ a "second of level of item similarity" that is based on the computed item similarity matrix and uses information on all items. Examples of such a second step is Euclidean distance or correlation. Similarity of items i and j is given by the Euclidean distance or Pearson correlation of rows i and j in the similarity matrix. Note that Euclidean distance may be used implicitly when we use standard implementation of some clustering algorithms (e.g., k-means).

With the basic approach to item similarity, we consider items similar when performance of learners on these items is similar. With the second step of item similarity, we consider two items similar when they behave similarly with respect to other items. The main reason for using this second step is the reduction of noise in data by using more information. This may be useful particularly to deal with learning. Two very similar items may have rather low direct similarity, because getting a feedback on the first item can strongly influence the performance on the second item. However, we expect both items to have similar similarities to other items.

A more technical reason to using the second step (particularly the Euclidean distance) is to obtain a measure that is a distance metric. The measures described above mostly do not satisfy triangle inequality and thus do not satisfy the requirements on distance metric; this property may be important for some clustering algorithms.

3. EVALUATION

In this work we focus on item similarity, but we keep the overall context depicted in Figure 1 in mind. The quality of a visualization is to a certain degree subjective and difficult to quantify, but the quality of clusters can be quantified and thus we can use it to compare similarity measures. From the large pool of existing clustering algorithms [15] we consider k-means, which is the most common implementation of centroid-based clustering, and hierarchical clustering. We used agglomerative or "bottom up" approach where items are successively merged to clusters using Ward's method as linkage criteria.

3.1 Data

We use data from real educational systems as well as simulated learner data. Real-world data provide information about the realistic performance of techniques, but the evaluation is complicated by the fact that we do not know the "ground truth" (the "correct" similarity or clusters of items). Simulated data provide a setting that is in many aspects simplified but allows easier evaluation thanks to the access to the ground truth.

For generating simulated data we use a simple approach with minimal number of assumptions and ad hoc parameters. Each item belongs to one of k knowledge components. Each knowledge component contains n items. Each item has a difficulty generated from the standard normal distribution $d_i \sim \mathcal{N}(0, 1)$. Skills of learners with respect to individual knowledge components are independent. Skill of a learner l with respect to knowledge component j is generated from the standard normal distribution $\theta_{lj} \sim \mathcal{N}(0, 1)$. We assume no learning (constant skills). Answers are generated as Bernoulli trials with the probability of a correct answer given by the logistic function of the difference of a relevant skill and an item difficulty (a Rasch model): $p = exp(\theta_{lj} - d_i)^{-1}$. This approach is rather standard, for example Piech at al. [26] use very similar procedure and also other works use closely related procedures [4, 12]. In the experiment reported below the basic setting is 100 learners, 5 knowledge components with 20 items each.

Table 2: Data used for analysis.

	learners	items	answers
Czech 1 (adjectives)	1 134	108	62 613
Czech 2	4 567	210	336 382
MatMat: numbers	6 434	60	67 753
MatMat: addition	3 580	135	20 337
Math Garden: addition	83 297	30	881 994
Math Garden: multiplic.	97 842	30	1 233 024

To evaluate techniques on realistic educational data, we use data from three educational systems. Table 2 describes the size of the used data sets.

Umíme Česky (`umimecesky.cz`) is a system for practice of Czech spelling and grammar. We use data only from one exercise from the system – simple "fill-in-the-blank" questions with two options. We use only data on the correctness of answers (response time is available, but since it depends on the text of a particular item its utilization is difficult). We focus particularly on one subset of items: questions about the choice between i/y in suffixes of Czech adjectives. For this subset we have manually determined 7 groups of items corresponding to Czech grammar rules.

MatMat (`matmat.cz`) is a system for practice of basic arithmetic (e.g., counting, addition, multiplication). For each item we know the underlying construct (e.g., "13" or "7 + 8") and also the specific form of questions (e.g., what type of visualization has been used). We use data on both correctness and response time. We selected the two largest subsets: multiplication and numbers (practice of number sens, counting).

Math Garden is another system for practice of basic arithmetic [16]. This system is more widely used than MatMat, but we do not have direct access to the system and detailed data. For the analysis we reuse publicly available data from previous research [6]. The available data contain both correctness of answers and response times, but they contain information only about 30 items without any identification of these items.

3.2 Comparison of Similarity Measures

To evaluate similarity measures we consider several types of analysis. With simulated data, we analyze the similarity measures with respect to the ground truth while for real-world data we evaluate correlations among similarity measures. We also compare the quality of subsequent clusterings using adjusted Rand index (ARI) [27, 31], which measures the agreement of two clusterings (with a correction for agreement due to chance). Typically, we use the adjusted Rand index to compare the clustering with a ground truth (available for simulated data) or with a manually provided

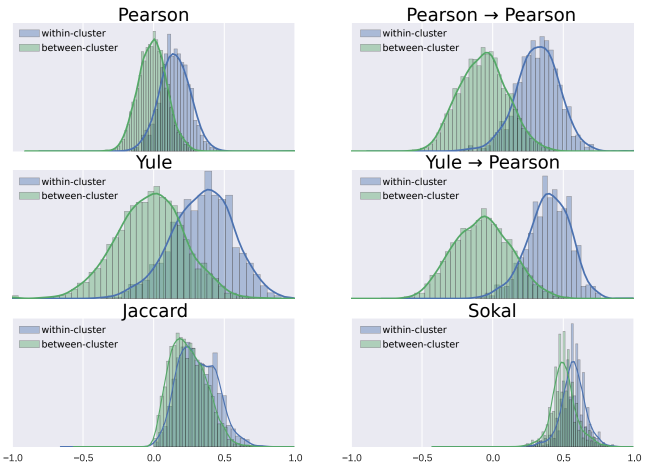

Figure 2: Differences between similarity values inside knowledge components and between them. Simulated data set with the basic setting were used.

classification (available for the Czech 1 data set). It can be also used to compare two detected clusterings (clusterings based on two different algorithms or clusterings based on two independent halves of data).

As a first step in the evaluation of similarity measures, we consider experiments with simulated data where we can utilize the ground truth. In clustering we expect high within-cluster similarity values and low between-cluster similarity values. Figure 2 shows distribution of the similarity values for selected measures and suggest which measures separate within-cluster and between-cluster values better and therefore which measures will be more useful in clustering. The results show that for Jaccard and Sokal measures the values overlap to a large degree, whereas Pearson and Yule measures provide better results. Adding the second step – Pearson correlation in this example – to the similarity measure separates within-cluster and between-cluster values better. That suggests that extending similarities in this way is not only necessary step for some subsequent algorithms such as k-means but also a useful technique with better performance.

For data coming from real systems we do not know the ground truth and thus we can only compare the similarity measures to each other. To evaluate how similar two measures are we take all similarity values for all item pairs and computed correlation coefficient. Figure 3 shows results for two data sets which are good representatives of overall results. Pearson and Cohen measures are highly correlated (> 0.98) across all data sets and have nearly the same values (although not exactly the same). Larger differences (but only up to 0.1) can be found typically when one of the values in the agreement matrix is small and that happens only for poorly correlated items with the resulting similarity value around 0. The second pair of highly correlated measures is Ochiai and Jaccard, which are both asymmetric with respect to the agreement matrix. The correlation between these two pairs of measures vary depending on data set and in some cases drops up to 0.5. Because of this high correlation within these pairs we further report results only for Pearson and Jaccard measures. Yule measure is usually similar to Pearson measure (correlation usually around 0.9). The main difference is that the Yule measure spreads values more evenly across the interval [-1, 1]. Sokal is the most outlying measure with no correlation or small correlation (usually < 0.6) with all other measures.

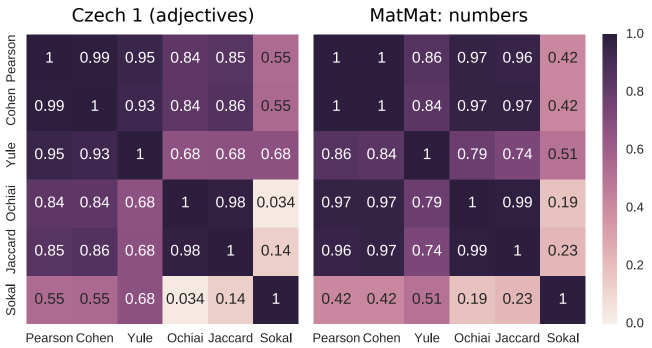

Figure 3: Correlations of similarity measures.

Figure 4 shows the effect of the second levels of item similarity on the Pearson measure (results for other measures are analogical). The Euclid distance as second level similarity brings larger differences (lower correlation) than Pearson correlation. The correlations for large data sets such as Math Garden are usually high (> 0.9) and conversely the lowest correlations are found in results for small data sets. This suggests that the second level of similarity is more significant, and thus potentially more useful, where only limited amount of data is available.

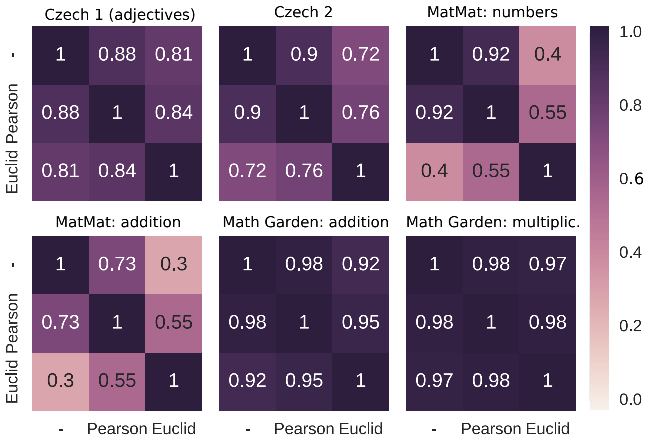

Figure 4: Correlations of Pearson measure and Pearson with different second levels.

Finally, we evaluate the quality of the similarity measures according to the performance of the subsequent clustering. From the two considered clustering methods we used the hierarchical clustering in this comparison because it naturally works with similarity measure and does not require metric space. The other two methods have similar result with same conclusions. Table 3 and Figure 5 show results. Although the results are dependent on the specific data set and the used clustering algorithm, there is quite clear general conclusion. Pearson and Yule measures provide better results

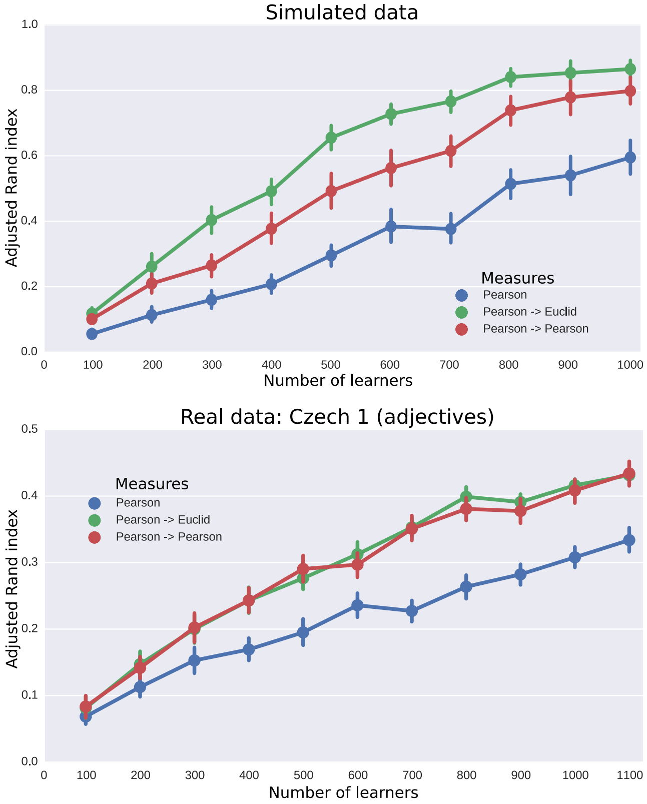

Figure 5: The quality of clustering for different measures used in the second step of item similarity. Top: Simulated data with 5 correlated skills. Bottom: Czech grammar with 7 manually determined clusters.

than Jaccard and Sokal, i.e., for the considered task the later two measures are not suitable. The Pearson is usually slightly better than Yule but the choice between them seems not to be fundamental (which is not surprising given that they are highly correlated). The results also show that the "second step" is always useful. The result for simulated data favor Euclidean distance over Pearson but there are almost no differences for real-world data.

3.3 Do We Have Enough Data?

In machine learning the amount of available data often is more important than the choice of a specific algorithm [2]. Our results suggest that once we choose a suitable type of similarity measure (e.g., Pearson, Cohen, or Yule), the differences between these measures are not fundamental, the more important issue becomes the size of available data.

Specifically, for a given data set we want to know whether the data are sufficiently large so that the computed item similarities are meaningful and stable. This issue can be explored by analyzing confidence intervals for computed similarity values. As a simple approach to analysis of similarity stability we propose the following approach: We split the available data into two independent halves (in a learner stratified manner), for each half we compute the item similarities, and we compute the correlation of the resulting item similarities.

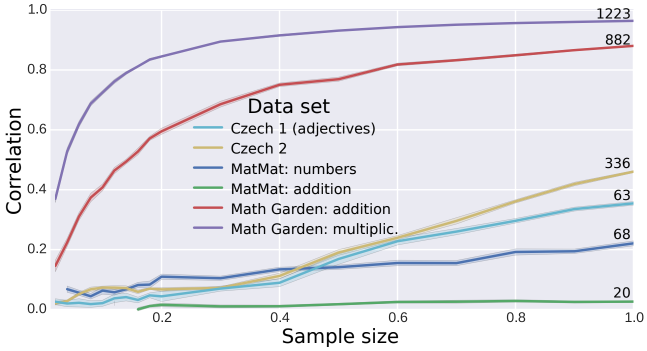

Figure 6: Stability of similarity measure (Yule) for real-world data sets. Data set was sampled, split to halves and Pearson correlation was computed for similarity values. Numbers on the right side indicate thousands of answers in data sets.

We can also perform this computation for artificially reduced data sets – this shows how the stability of results increases with the size of data. Figure 6 shows this kind of analysis for our data (real-world data sets). We clearly see large differences among individual data sets. Math Garden data set contains large number of answers and only a few items, the results show excellent stability, clearly in this case we have enough data to analyze item similarities. For the Czech grammar data set we have large number of answers, but these are divided among relatively large number of items. The results show a reasonably good stability, the data are usable for analysis, but clearly more data can bring improvement. For MatMat data the stability is poor, to draw solid conclusions about item similarities we need more data.

3.4 Response Time Utilization

The incorporation of response time information to similarity measure can change the meaning of similarity. Figure 7 gives such example and shows projection of items from MatMat practicing number sense. Similar items according to measures using only correctness of answers tend to be items with the same graphical representation in the system. On the other hand, similar items according to measures using also response time are usually items practicing close numbers.

We used this method also on data sets from Math Garden, which are much larger. In this case the use of response times has only small impact on the computed item similarities (correlations between 0.9 and 0.95). However, the use of response times influences how quickly does the computation converge, i.e., how much data do we need. To explore this we consider as the ground truth the average of computed similarity matrices with and without response times for the whole data set. Then we used smaller samples of the data set, used them to compute item similarities and checked the agreement with this ground truth. Figure 8 shows the difference between speed of convergence of measure with and without response time utilization. Results shows that the measure which use addition information from response time converges to ground truth much faster. This result suggests that the use of response time can improve clustering or visualizations when only small number of answers are available.

Table 3: Comparison of similarity measures for one real-world data (with sampled students) set and simulated data sets with c knowledge components and l learners. The values provide the adjusted Rand index (with 0.95 confidence interval) for a hierarchical clustering computed based on the specific similarity measure. The top result for every data set is highlighted.

	Czech 1 (c=7)	$l = 50, c = 5$	$l = 100, c = 5$	$l = 200, c = 5$	$l = 100, c = 2$	$l = 100, c = 10$
Pearson	0.32 ± 0.02	0.26 ± 0.04	0.48 ± 0.05	0.84 ± 0.05	0.77 ± 0.12	0.34 ± 0.04
Jaccard	0.31 ± 0.03	0.06 ± 0.03	0.15 ± 0.04	0.29 ± 0.08	0.32 ± 0.18	0.09 ± 0.02
Yule	0.31 ± 0.03	0.19 ± 0.04	0.43 ± 0.05	0.77 ± 0.07	0.60 ± 0.15	0.31 ± 0.03
Sokal	0.15 ± 0.06	0.11 ± 0.02	0.18 ± 0.03	0.25 ± 0.05	0.12 ± 0.11	0.14 ± 0.02
Pearson → Euclid	$\mathbf{0.43} \pm 0.01$	$\mathbf{0.45} \pm 0.05$	$\mathbf{0.80} \pm 0.06$	$\mathbf{0.98} \pm 0.01$	$\mathbf{0.95} \pm 0.03$	$\mathbf{0.67} \pm 0.04$
Yule → Euclid	0.32 ± 0.02	0.36 ± 0.05	0.65 ± 0.07	0.94 ± 0.04	0.89 ± 0.11	0.43 ± 0.03
Pearson → Pearson	0.41 ± 0.03	0.39 ± 0.05	0.73 ± 0.06	0.96 ± 0.02	0.92 ± 0.03	0.55 ± 0.04
Yule → Pearson	0.32 ± 0.03	0.38 ± 0.05	0.72 ± 0.06	0.97 ± 0.02	0.94 ± 0.04	0.55 ± 0.05

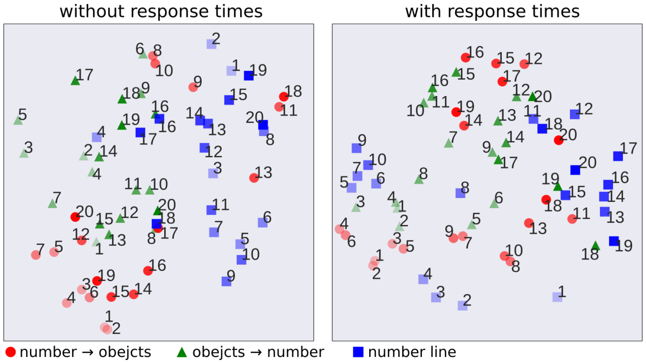

Figure 7: Projection of items practicing number sense from MatMat system. Left: Measure based only correctness. Right: Measure using response time. Opacity corresponds to the number value of the item and color corresponds to the graphical representation of the task.

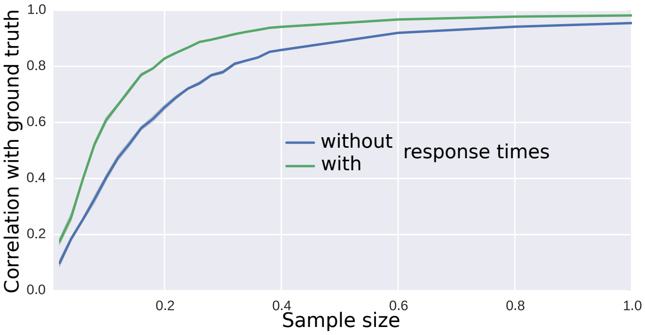

Figure 8: The speed of convergence to ground truth for measures with and without response time on Math Garden addition data set.

4. DISCUSSION

Our focus is the automatic computation of item similarities based on learners' performance data. These similarities can be then used in further analysis of an item relations such as an item clustering or a visualization. This outlines direction for future work in which methods using the item similarities should be studied in more detail. Compared to alternative approaches that have been proposed for the task (e.g., matrix factorizations, neural networks), the item similarity approach is rather straightforward, easy to realize, and it can be easily combined with other sources of information about items (text of items, expert opinion). For these reasons the item similarity approach should be used at least as a baseline in proposals for more complex methods like deep knowledge tracing [26].

The most difficult step in this approach is the choice of a similarity measure. Once we make a specific choice, the realization of the approach is easy. Our results provide some guidelines for this choice. Pearson, Yule, and Cohen measures lead to significantly better results than Ochiai, Sokal, and Jaccard measures. It is also beneficial to use the second step of item similarity (e.g., the Euclidean distance over vectors of item similarities). The exact choice of details does not seem to make fundamental difference (e.g., Pearson versus Yule in the first step, the Euclidean distance versus Pearson correlation in the second step). The Pearson correlation coefficient is a good "default choice", since it provides quite robust results and is applicable in several settings and steps. It also has the pragmatic advantage of having fast, readily available implementation in nearly all computational environments, whereas measures like Yule may require additional implementation effort.

The amount of data available is the critical factor for the success of automatic analysis of item relations. A key question for practical applications is thus: "Do we have enough data to use automated techniques?" In this work we used several specific methods for analysis of this question, but the issue requires more attention – not just for the item similarity approach, but also for other methods proposed in previous work. For example previous work on deep knowledge tracing [26], which studies closely related issues, states only that deep neural networks require large data without providing any specific quantification what 'large' means. The necessary quantity of data is, of course, connected to the quality of data – some data sources are more noisy than other, e.g., answers from voluntary practice contain more noise than answers from high-stakes testing. An important direction for future work is thus to compare model based and item similarity approaches while taking into account the ‘amount and quality of data available’ issue.

5. REFERENCES

[1] R. S. Baker. Stupid tutoring systems, intelligent humans. *International Journal of Artificial Intelligence in Education*, 26(2):600–614, 2016.

[2] M. Banko and E. Brill. Scaling to very very large corpora for natural language disambiguation. In *Proc. of Association for Computational Linguistics*, pages 26–33, 2001.

[3] T. Barnes. The q-matrix method: Mining student response data for knowledge. In *Educational Data Mining Workshop*, 2005.

[4] W.-H. Chen and D. Thissen. Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, 22(3):265–289, 1997.

[5] S.-S. Choi, S.-H. Cha, and C. C. Tappert. A survey of binary similarity and distance measures. *Journal of Systemics, Cybernetics and Informatics*, 8(1):43–48, 2010.

[6] F. Coomans, A. Hofman, M. Brinkhuis, H. L. van der Maas, and G. Maris. Distinguishing fast and slow processes in accuracy-response time data. *PloS one*, 11(5):e0155149, 2016.

[7] M. Deshpande and G. Karypis. Item-based top-n recommendation algorithms. *ACM Transactions on Information Systems (TOIS)*, 22(1):143–177, 2004.

[8] M. C. Desmarais. Mapping question items to skills with non-negative matrix factorization. *ACM SIGKDD Explorations Newsletter*, 13(2):30–36, 2012.

[9] M. C. Desmarais and R. S. Baker. A review of recent advances in learner and skill modeling in intelligent learning environments. *User Modeling and User-Adapted Interaction*, 22(1-2):9–38, 2012.

[10] M. C. Desmarais, B. Beheshti, and P. Xu. The refinement of a q-matrix: Assessing methods to validate tasks to skills mapping. In *Proc. of Educational Data Mining*, pages 308–311, 2014.

[11] C. Desrosiers and G. Karypis. A comprehensive survey of neighborhood-based recommendation methods. In *Recommender systems handbook*, pages 107–144. Springer, 2011.

[12] H. Finch. Comparison of distance measures in cluster analysis with dichotomous data. *Journal of Data Science*, 3(1):85–100, 2005.

[13] J. Friedman, T. Hastie, and R. Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics Springer, Berlin, 2001.

[14] D. A. Jackson, K. M. Somers, and H. H. Harvey. Similarity coefficients: measures of co-occurrence and association or simply measures of occurrence? *American Naturalist*, pages 436–453, 1989.

[15] A. K. Jain. Data clustering: 50 years beyond k-means. *Pattern recognition letters*, 31(8):651–666, 2010.

[16] S. Klinkenberg, M. Straatemeier, and H. Van der Maas. Computer adaptive practice of maths ability using a new item response model for on the fly ability and difficulty estimation. *Computers & Education*, 57(2):1813–1824, 2011.

[17] K. R. Koedinger, A. T. Corbett, and C. Perfetti. The knowledge-learning-instruction framework: Bridging the science-practice chasm to enhance robust student learning. *Cognitive science*, 36(5):757–798, 2012.

[18] Y. Koren and R. Bell. Advances in collaborative filtering. *Recommender Systems Handbook*, pages 145–186, 2011.

[19] A. S. Lan, A. E. Waters, C. Studer, and R. G. Baraniuk. Sparse factor analysis for learning and content analytics. *Journal of Machine Learning Research*, 15(1):1959–2008, 2014.

[20] S.-F. M. Liang and L.-W. Tzeng. Assessing suitability of similarity coefficients in measuring human mental models. In *Network of Ergonomics Societies Conference*, pages 1–5. IEEE, 2012.

[21] J. Nižnan, R. Pelánek, and J. Řihák. Using problem solving times and expert opinion to detect skills. In *Proc. of Educational Data Mining*, pages 434–434, 2014.

[22] J. Nižnan, R. Pelánek, and J. Řihák. Student models for prior knowledge estimation. In *Proc. of Educational Data Mining*, pages 109–116, 2015.

[23] M. O’Connor and J. Herlocker. Clustering items for collaborative filtering. In *Proc. of the ACM SIGIR Workshop on Recommender Systems*, volume 128. UC Berkeley, 1999.

[24] Y.-J. Park and A. Tuzhilin. The long tail of recommender systems and how to leverage it. In *Proc. of Recommender systems*, pages 11–18. ACM, 2008.

[25] R. Pelánek and J. Řihák. Properties and applications of wrong answers in online educational systems. In *Proc. of Educational Data Mining*, 2016.

[26] C. Piech, J. Bassen, J. Huang, S. Ganguli, M. Sahami, L. J. Guibas, and J. Sohl-Dickstein. Deep knowledge tracing. In *Advances in Neural Information Processing Systems*, pages 505–513, 2015.

[27] W. M. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336):846–850, 1971.

[28] J. Rihák. Use of time information in models behind adaptive system for building fluency in mathematics. In *Proc. of Educational Data Mining*, 2015.

[29] B. M. Sarwar, G. Karypis, J. Konstan, and J. Riedl. Recommender systems for large-scale e-commerce: Scalable neighborhood formation using clustering. In *Proc. of Computer and Information Technology*, volume 1, 2002.

[30] E. Şenyürek and H. Polat. Effects of binary similarity measures on top-n recommendations. *Anadolu University Journal of Science and Technology – A Applied Sciences and Engineering*, 14(1):55–65, 2013.

[31] N. X. Vinh, J. Epps, and J. Bailey. Information theoretic measures for clusterings comparison: is a correction for chance necessary? In *Proc. of Machine Learning*, pages 1073–1080. ACM, 2009.

[32] M. J. Warrens. On association coefficients for 2× 2 tables and properties that do not depend on the marginal distributions. *Psychometrika*, 73(4):777–789, 2008.