Data Integration through Clustering and Finding Statistical Relations - Validation of Approach

Marek Jaszuk, Teresa Mroczek, and Barbara Fryc

University of Information Technology and Management, ul. Sucharskiego 2, Rzeszow, Poland,
{mjaszuk,tmroczek,bfryc}@wsiz.rzeszow.pl

Abstract. The paper analyzes an approach to data integration based on finding statistical relations between data. The data used for experimenting comes from surveys collected from student groups. The practical problem that underlies this research is discovering the model of knowledge about students, which would allow for making predictions about their future educational success or failure. The obstacle is that data collected for different groups over years has different format and this makes it difficult to reuse the previously collected data. Thus we had to find a way, to overcome this difficulty and integrate the heterogeneous data. The paper analyzes the feasibility of integrating data using this method. Although based on particular application, the model of computations presented in the paper is of more general nature, and should be applicable in many other domains.

Key words: data integration, clustering, semantic class, correlation, survey

1 Introduction

Increase in popularity of information technologies brought us a large number of independently created and managed information systems. Such systems can contain similar information but coming from disparate sources which leads to information heterogeneity. This prevents from interoperability of such systems and their integration. Thus it is highly demanded to overcome the heterogeneities through some data integration technique, and data integration is one of the central problems in information systems.

Usually information heterogeneity can be considered on three levels: syntactic, structural and semantic. Our focus in this paper is on semantic heterogeneity and data integration on this level. The problem of semantic heterogeneity has been studied intensively in the past years (see e.g. [2, 3, 5, 6, 9]). Automatic identification of semantic relations between different data sets has been investigated [2] together with representation and using of identified relations for transferring data and query answering [1, 5, 9]. A prominent part of research is devoted to investigating the role of ontologies, which represent formally the conceptual structure of a given application domain. The ontologies are used for identifying and using semantic relations, necessary for representing information systems to be integrated. In this regard, an ontology works as an intermediary between heterogeneous data sources.

The main difficulty with using ontologies is that they are usually handcrafted by domain experts accompanied by ontology engineers. Any modification in an ontology requires human effort, and thus is inefficient. Another problem is that an ontology for a given domain can be developed in many different ways, and its optimal structure depends on particular application. Thus a particular ontology is not always suitable for a specific problem to be solved. In consequence, a method which automatizes integration of data coming from different sources, is highly desired.

The application example discussed in this paper is integration of different versions of survey data. Solving this problem using standard ontology based techniques would require developing ontologies for the survey questions, and then matching the ontologies to integrate the data. The semantic space for such an ontology is huge, because changing even a single word in a survey question can change the interpretation of this question, and thus shifts the meaning of this question. Thus doing the task of data integration manually can be a challenge. We propose an alternative approach based on finding statistical relations between data. This technique has been applied to surveys made on students, but it is more general, and could be applied to many other kinds of data with similar structure.

The paper is organized as follows. Sec. 2 discusses the practical problem to be solved. In Sec. 3 we discuss, how the survey questions are formulated. In Sec. 4 the basic assumptions about statistical data representation and the definition of semantic space. Then, in Sec. 5, we demonstrate the correlations between two data sets obtained independently, in order to confirm validity of the presented approach. Finally in Sec. 6 we discuss the way of integrating data coming from different surveys using clustering techniques.

2 The Problem Formulation

The practical problem that underlies considerations presented in this paper, is integration of data coming from groups of students. The University of Information Technology and Management in Rzeszów (UITM), where we conduct the research, collects some basic data about the students in the computer system, like the date of birth, the gender, the grades, etc. However, these data are not sufficient when one wants to perform more sophisticated kind of reasoning about the students. In our case, the main interest are in the future study results (educational success) of the students, that are beginning their education. Such information is interesting both for the group as a whole, as well as for selected individuals. The potential value of such information is both for the university authorities, as well, as for the teaching stuff, because it allows for early identification of potential problems, or outstanding individuals, who require special treatment. One of the factors that indicate the potential success are the results in the preceding stage of education, like the secondary school. This is, however, not a complete information, and the educational success is influenced by many other factors. We assume that the missing information about the factors influencing educational success can be collected by carrying out surveys, with questions related to socio-economic situation of students, as well as their motivations and reasons to study. The problem is in itself interesting from the social sciences perspective, but this aspect will not be discussed here.

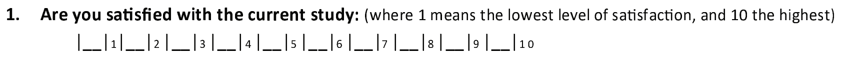
There is potentially a large number of details, that could be asked in such a survey. The problem is thus selection of questions to be included in the survey. Unfortunately, the survey cannot be to long, otherwise, the students would not be willing to fill it.

Thus the choice of the questions should be very careful. The most adequate collection of questions can be obtained through a trial and error method. Each survey has to be followed by statistical analysis, which would indicate the questions well correlated with the educational success categories. After several iterations we should be able collect the questions delivering desirable information. But even if we identify the questions, there is always a possibility, that in the future someone would like to incorporate some new questions in the survey. The reason for such a modification would be changes in the external situation, and identification of possible new factors, that could be relevant. All the well known reasoning methods are based on unified data sets, i.e. to make predictions for a new data set, this set should be composed of data in the same format, on which the reasoning machine has been trained. Any modification in the data, like introducing new questions, requires retraining of the reasoning machine on the new data. At the same time, all the previous data which are incompatible with the new format become useless. This is an important problem, because in this way we loose a huge amount of unique data, which were difficult to collect.

If one wants to avoid loosing the previously recorded data, it is necessary to match the data formats in some way. This is usually done through manual effort. One of the standard approaches to the problem is based on using semantic models (ontologies), which allow for integrating both the data sets through ontology mapping [9]. The difficulties related to using ontologies, have already been mentioned, and we want to avoid the direct ontology creation and mapping. The approach demonstrated in this paper tries to complete the task data integration by generating a set of classes for data automatically. The foundations for this method were described in [4].

3 The Surveys

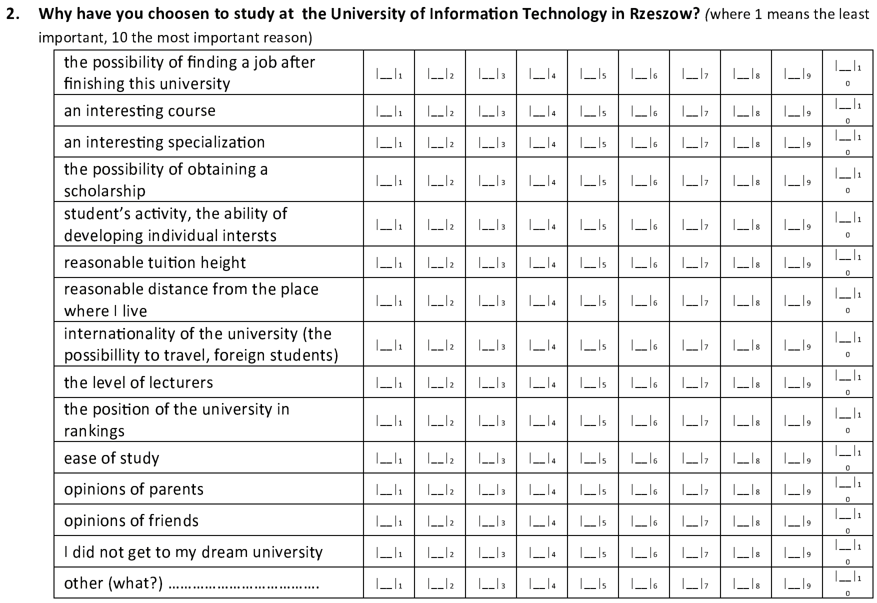
The questions for the surveys were prepared by social sciences experts according to their best knowledge. They were not tailored specifically for our experiments, but just to collect the data about the students, like this is done for other kinds of investigation. The first survey, that we analyzed, contained 21 questions with different structure depending on the question specifics. The structure of the questions was organized to make them clear and understandable. Fig. 1 shows the first question, which is a choice between 10 values from the range 1-10, reflecting the opinion of the questioned person. Some of the questions were actually groups of the questions related to the same subject (Fig. 2). Thus the actual number of questions is much larger than 21, due to subquestions. Yet another type of questions are multiple choice questions (Fig. 3)



1. **Are you satisfied with the current study:** (where 1 means the lowest level of satisfaction, and 10 the highest)

|__|1|__|2 |__|3 |__|4 |__|5 |__|6 |__|7 |__|8 |__|9 |__|10

Fig. 1. The question with choice between 10 values



2. **Why have you choosen to study at the University of Information Technology in Rzeszow?** *(*where 1 means the least important, 10 the most important reason)

<table>
<tr><td>the possibility of finding a job after finishing this university</td><td>|__|1</td><td>|__|2</td><td>|__|3</td><td>|__|4</td><td>|__|5</td><td>|__|6</td><td>|__|7</td><td>|__|8</td><td>|__|9</td><td>|__|10</td></tr>
<tr><td>an interesting course</td><td>|__|1</td><td>|__|2</td><td>|__|3</td><td>|__|4</td><td>|__|5</td><td>|__|6</td><td>|__|7</td><td>|__|8</td><td>|__|9</td><td>|__|10</td></tr>
<tr><td>an interesting specialization</td><td>|__|1</td><td>|__|2</td><td>|__|3</td><td>|__|4</td><td>|__|5</td><td>|__|6</td><td>|__|7</td><td>|__|8</td><td>|__|9</td><td>|__|10</td></tr>
<tr><td>the possibility of obtaining a scholarship</td><td>|__|1</td><td>|__|2</td><td>|__|3</td><td>|__|4</td><td>|__|5</td><td>|__|6</td><td>|__|7</td><td>|__|8</td><td>|__|9</td><td>|__|10</td></tr>
<tr><td>student's activity, the ability of developing individual intersts</td><td>|__|1</td><td>|__|2</td><td>|__|3</td><td>|__|4</td><td>|__|5</td><td>|__|6</td><td>|__|7</td><td>|__|8</td><td>|__|9</td><td>|__|10</td></tr>
<tr><td>reasonable tuition height</td><td>|__|1</td><td>|__|2</td><td>|__|3</td><td>|__|4</td><td>|__|5</td><td>|__|6</td><td>|__|7</td><td>|__|8</td><td>|__|9</td><td>|__|10</td></tr>
<tr><td>reasonable distance from the place where I live</td><td>|__|1</td><td>|__|2</td><td>|__|3</td><td>|__|4</td><td>|__|5</td><td>|__|6</td><td>|__|7</td><td>|__|8</td><td>|__|9</td><td>|__|10</td></tr>
<tr><td>internationality of the university (the possibillity to travel, foreign students)</td><td>|__|1</td><td>|__|2</td><td>|__|3</td><td>|__|4</td><td>|__|5</td><td>|__|6</td><td>|__|7</td><td>|__|8</td><td>|__|9</td><td>|__|10</td></tr>
<tr><td>the level of lecturers</td><td>|__|1</td><td>|__|2</td><td>|__|3</td><td>|__|4</td><td>|__|5</td><td>|__|6</td><td>|__|7</td><td>|__|8</td><td>|__|9</td><td>|__|10</td></tr>
<tr><td>the position of the university in rankings</td><td>|__|1</td><td>|__|2</td><td>|__|3</td><td>|__|4</td><td>|__|5</td><td>|__|6</td><td>|__|7</td><td>|__|8</td><td>|__|9</td><td>|__|10</td></tr>
<tr><td>ease of study</td><td>|__|1</td><td>|__|2</td><td>|__|3</td><td>|__|4</td><td>|__|5</td><td>|__|6</td><td>|__|7</td><td>|__|8</td><td>|__|9</td><td>|__|10</td></tr>
<tr><td>opinions of parents</td><td>|__|1</td><td>|__|2</td><td>|__|3</td><td>|__|4</td><td>|__|5</td><td>|__|6</td><td>|__|7</td><td>|__|8</td><td>|__|9</td><td>|__|10</td></tr>
<tr><td>opinions of friends</td><td>|__|1</td><td>|__|2</td><td>|__|3</td><td>|__|4</td><td>|__|5</td><td>|__|6</td><td>|__|7</td><td>|__|8</td><td>|__|9</td><td>|__|10</td></tr>
<tr><td>I did not get to my dream university</td><td>|__|1</td><td>|__|2</td><td>|__|3</td><td>|__|4</td><td>|__|5</td><td>|__|6</td><td>|__|7</td><td>|__|8</td><td>|__|9</td><td>|__|10</td></tr>
<tr><td>other (what?) ………………………………</td><td>|__|1</td><td>|__|2</td><td>|__|3</td><td>|__|4</td><td>|__|5</td><td>|__|6</td><td>|__|7</td><td>|__|8</td><td>|__|9</td><td>|__|10</td></tr>
</table>

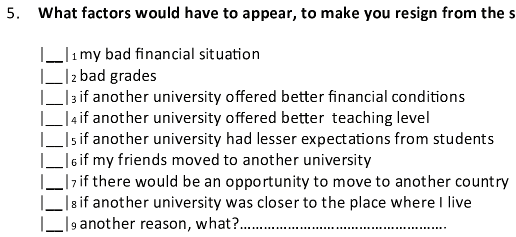
Fig. 2. The question which is a combination of multiple subquestions

4 The Data Model

4.1 The Survey Representation

As we can see, the questions are of different form, and have to be reduced into a homogeneous format to allow for treating them in the same way. We do this by separating each possible outcome of the questions (an answer), and treating it as a separate attribute. In this way, for the first question in Fig. 1, we get 10 different question/answer pairs i.e:

- 1. Are you satisfied with the current study - 1
- 2. Are you satisfied with the current study - 2
- ...



5. **What factors would have to appear, to make you resign from the study?** (choose not more than 3 factors)

|__|1 my bad financial situation
|__|2 bad grades
|__|3 if another university offered better financial conditions
|__|4 if another university offered better teaching level
|__|5 if another university had lesser expectations from students
|__|6 if my friends moved to another university
|__|7 if there would be an opportunity to move to another country
|__|8 if another university was closer to the place where I live
|__|9 another reason, what?……………………………………………

Fig. 3. The multiple choice question

- 20. Are you satisfied with the current study - 10

The rest of the questions are decomposed in the same way. In consequence the initial number of 21 questions transforms into the space of more than 500 question/answer pairs. The survey after completing by a student is a binary vector:

$$O_i = \{I_k^i : k \in 1, \dots M\}, \tag{1}$$

where O_i is the vector representing i-th surveyed student, I_k^i is the k-th coordinate of i-th survey vector. M is the number of possible question/answer pairs in the survey. The vector contains 1-s in positions representing the answers selected by a student, and 0-s for answers which were not selected.

4.2 Educational Success Categories

Except the surveys, we collected the data about results of education for each of the students that filled the survey. These data are available in the university computer system. The results of education are the grades, or the information that the study has been broken for some reason. These data is available not earlier, than after the end of the first semester, while the survey was completed in the beginning of the academic year. The survey data are collected in the first semester of the university course. In this way we are able to follow the results of the students from the beginning till the end of the study, and confront them with the survey answers.

Our method requires dividing the investigated students into a number of groups related to their educational success. We do this by applying hierarchic clustering [7][1] within the space of grades that the students got. The applied method of clustering was chosen because it allows for generating different numbers of clusters, depending on the choice of the cut point on the hierarchy. We performed validation of a number of well known clustering algorithms, and most of them revealed comparable clustering quality. So this factor was not crucial for the choice of clustering method.

We had no a priori assumption about the number of clusters, so we decided to choose the cut point that generated 5 clusters. This number seemed suitable to our experiments, although we do not exclude the possibility of experimenting also with other numbers of clusters. There was also a number of students, who filled the survey, but had no grades, because their study had been broken. This class of students are of particular interest, because they represent the educational failure. In consequence we got 6 categories of students - 5 coming from clustering, and one of those, who resigned from study.

4.3 The Semantic Distance

Our purpose is integration of the survey data. To be able to integrate them automatically, we have to start from determining the semantic distance between survey answers. The context data, which allow for determining the distance, are the educational success

[1] Implemented in the R Project [8]

categories. The measure of the distance is based on statistical distribution of the survey answers with respect to the context:

$$P_{I_k} = (P_{k1}, P_{k2}, \dots, P_{kN}), \tag{2}$$

where P_{I_k} is the frequency distribution vector of the survey answer I_k with respect to the N success categories (in our case there are 6 categories). P_{kn} is the frequency with which the answer numbered k (I_k) was found in the success category numbered n, $(n = 1, \dots, N)$, i.e. the chance that student belonging to success category n, chooses the answer k.

The space spanned by the distributions P_{I_k} plays the role of semantic space. The direction of the distribution vector (2) represents the meaning of every survey answer. The semantic distance between two answers is measured by the angle between respective distribution vectors. For practical reasons it is more convenient to use the cosine of the angle between the vectors. The cosine is the semantic similarity measure which ranges between 0 and 1. This range results from the frequencies, which are non negative values, and thus the angle between vectors never exceeds $\pi/2$. The answers with identical meaning have the maximal similarity equal to 1, and the answers with completely different meaning have similarity equal to 0. The semantic similarity S_{kl} between two answers I_k and I_l is calculated as:

$$S_{kl} = \cos\alpha_{kl} = \frac{P_{I_k} \cdot P_{I_l}}{\|P_{I_k}\| \, \|P_{I_l}\|}, \tag{3}$$

where α_{kl} is the angle between P_{I_k} and P_{I_l} vectors.

The justification for the thesis, that the semantic distance can be measured using (3) is the observation, that if there would be two questions in the survey, with identical meaning, they have to generate the same probability distribution (similarity=1). Otherwise that would would mean that surveyed persons interpreted the questions differently, and thus their meaning is different. The other possibility is that the survey was filled randomly, but we believe that this is not the case. It should be also noted that the similarity equal to 1 does not always mean that the human interpretation of the questions is identical. It is possible, that the interpretation is different, but still generates similar distribution. In terms of semantic model, this can be interpreted as synonymic question and answer. No mater which of the synonyms is used in the survey, the result is the same. So for computational purposes the synonymic questions are not a problem, because the reasoning based on them will be the same. A more detailed discussion of motivations for using the space of probabilities as the semantic space can be found in [4].

5 Correlations of Probability Distributions

The survey data are collected to build a computational model based on the groups of students that are currently studying, in order to be able to make predictions about the students recruited in the future. This approach makes sense only when the statistical distributions are stable for student groups from subsequent years. Thus it is necessary to verify the stability of distributions.

To assess the stability of the frequency distribution (2) for every answer in the survey, we conducted the survey for two student groups in subsequent years. The two surveys were not identical. They contained a number of questions that were identical, and a number of questions that were different. For the purpose of verification, only the common part of the questions is useful, so the rest of the question/answer pairs is not used for the stability verification purpose. Computing distributions (2) requires information about the grades of the surveyed students, to classify each of them into one of the 6 previously assumed categories of educational success. We clustered the first (older) of the investigated groups, and used the the same clustering for the second (younger) group. In this way the have a consistent classification system for students from both of the groups.

Given the classification we can compute the distributions (2) for each of the two groups. Then we can compare the distributions by computing the cosine of the angle between the distribution vectors for the same answers (semantic distance), but obtained from two different groups:

$$S_{kk}^{12} = \cos \alpha_{kk}^{12} = \frac{P_{I_k}^1 \cdot P_{I_k}^2}{\left\| P_{I_k}^1 \right\| \left\| P_{I_k}^2 \right\|}, \tag{4}$$

where P^1, P^2 are the frequency distributions obtained for group 1 and group 2 of the students respectively.

Ideally, the similarity (4) should be close to 1 for each of the answers. In reality this value spreads over the whole range of possible values (Fig. 4(a)). There is an observable number of answers, which correlate between the groups (semantic similarity close to 1). But there are also answers, which do not correlate at all.

We were able to identify the reason for low correlations easily - the answers with low correlation come from the answers rarely chosen by the students. Some of the answers were chosen by just a few persons, or even by no one. With such small numbers no reliable statistical distribution can be determined. Luckily one of the significant reasons for low numbers was easy to eliminate. This was the wide range of possible answers to particular questions - in many cases this range was set to 10 single choice values, like in questions in Figs. 1 and 2. The total number of students in each of the investigated groups was about 200, so statistically the average number of persons that should choose each of the answers should be 20. However, the students clearly preferred some answers than the others. Actually we anticipated this situation, and chosen such a wide range deliberately, because it is easy to reduce the range afterwards, in case if the original range did not work. So the range was reduced from 10 down to 5 possible choices. This reduced the number of the considered answers form 401 to 255, and immediately increased the number of students who chosen each of the answers. After computing the similarity between groups we got the similarity distribution presented in Fig. 4(b).

As it can be observed, the degree of highly correlated answers increased after reducing the range of answers (e.g. the number of answers with correlation higher than 0.9 increased from about 25% to about 50%). However, there are still answers which do not correlate well. Despite applying the answer reduction trick, there are still answers, which are less likely to be chosen by the students. According to our findings, this is the main reason for low correlations, which result from less statistically reliable distribu-

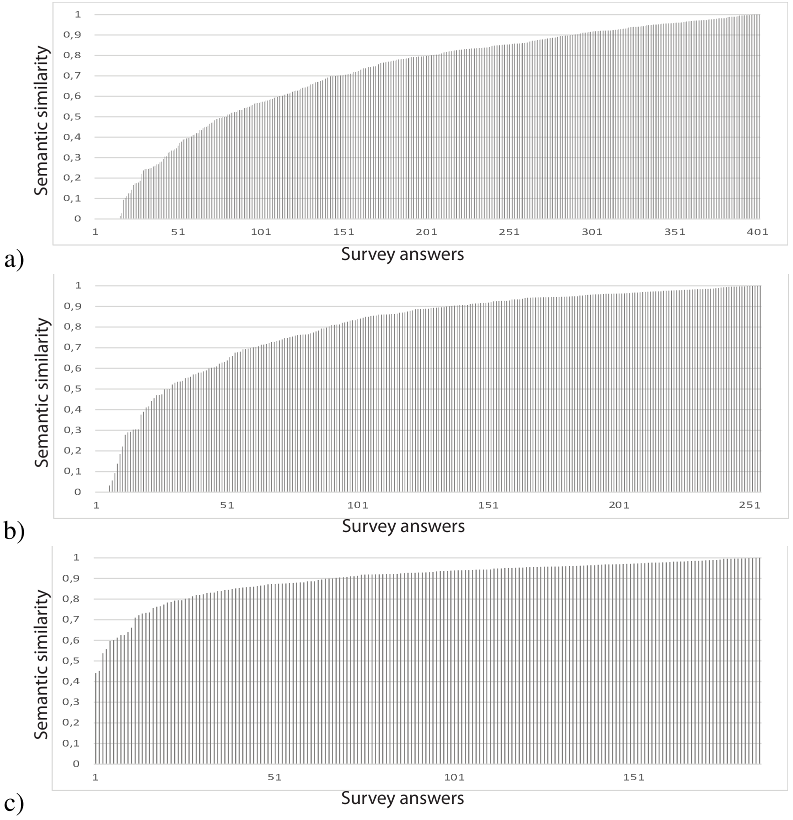


Fig. 4. Similarity distribution for the survey answers: (a) initial range of answers, (b) reduced range of answers, (c) further reduction of the range of answers

tions. There was still some space for reducing the range of answers, e.g. to 3 or even 2 possible choices, which can increase the level of correlations. Thus we applied further reduction of the range of answers to 3 possible choices (Fig. 4(c)). The result confirmed our suspections - the correlations further increased, and the number of uncorrelated answers vanished completely. The lowest similarity between the answers was on the level of 0.44.

Although the results look much better than the initial, there is still some space for improvements, which is related to several issues. The first of them are the survey questions. Not all them have the kind of structure which allow for their easy reduction. Thus there are still question/answer pairs, which are unlikely to be chosen by the students. The possible ways of increasing the correlations, would be:

1. eliminate the weakly correlated answers - the risk is, that in this way we will eliminate valuable information referring to a relatively small number of students,
2. reformulate the questions in order to force the students, to choose some answers more frequently - this is something that we consider to do in the future years,

3. increase the number of investigated students - we investigated students only from the Information technology specialization. Due to limited number of students, increasing the number would require extending the research onto other specializations, which is possible, but we are not sure if students, from very different specializations, will generate the same probability distributions. This is an interesting topic for the future research,
4. decrease the number of student success groups - the division into 6 groups might be too fine grained, thus we consider decreasing it to e.g. 4 groups, which immediately increases the number of students in each of the groups, and makes the distribution more reliable.

To summarize the results of correlation investigation, we can state that if the number of students, who selected a particular answer is sufficiently large, then the frequency distribution with respect to the educational success categories, in a vast number of cases remains stable. Thus such an answer can be used as a reliable indicator of the possible success category.

6 Data Integration through Clustering

As already mentioned, a survey can potentially contain many different questions, and along time some questions could be replaced by others. This makes it difficult to reuse the knowledge collected in the previous years, because the evolution of questions could lead to a potentially large set of questions. Finding relations among such questions along the timeline is a difficult task. Thus we develop the mechanism, which should allow for integrating the the old versions of surveys, with the newly created ones. The basis for this task is the already introduced semantic space of frequency distributions (2) together with the similarity measure (3).

The basic semantic relation, that can be discovered, among survey answers is the synonymy relation, i.e. finding answers with the same meaning. This task can be completed with the clustering technique. We applied the hierarchic clustering again due to its flexibility, and possibility of selecting various levels of clustering granularity. We used the cosine distance, to measure the distance between clustered objects (the survey answers), because this is the assumed semantic similarity measure. Here again it is interesting to assess, whether the closely related questions indeed fall in the same category.

There is a number of testing scenarios that can be proposed here. Because we want to integrate the data coming from surveys obtained from subsequent years, the best approach is to check, whether the answers belonging to some cluster for one year, belong to the same cluster in the subsequent year. To verify this, we clustered the answers for the first of the surveyed groups. Then we calculated the frequency distribution (2) for each of the answers in the second of the surveyed groups. This allows for determining the cluster (obtained on the first group), that each of the answers collected in the second group belongs to.

In the ideal situation, all of the answers for the second group should belong to the same clusters as for the first group. The results revealed that the situation is more complex to analyze. First of all, there are huge differences in the number of objects in each of the clusters. That of course depends on the cut level in the hierarchy. But in general, the majority of answers are grouped in several huge clusters. This is illustrated in Fig. 5 for the cut level in the clustering hierarchy equal to 0.1 (the parameter ranges between 0 and 1 - the lower the value, the larger number of clusters). The largest cluster contained 54 answers. The second group are the middle sized clusters (2 to 5), where the number of answers ranged between 31 and 12. The third group are the smallest clusters (6 to 14), where the number of answers ranged between 7 and 1. This is an interesting result, because it brings us insight into the nature of the gathered information. We can see, that the answers grouped in the huge clusters, do not bring much new information. In fact, we could resign from using all the question/answer pairs belonging to such clusters, and leave just one of them for each of the clusters. This would reduce the survey complexity significantly. More interesting are the answers grouped in small clusters. Their uniqueness indicate, that they bring some valuable information about the students, which distinguishes them from the others. This also indicates the possible regions, in which the survey could be extended to gather more useful information.

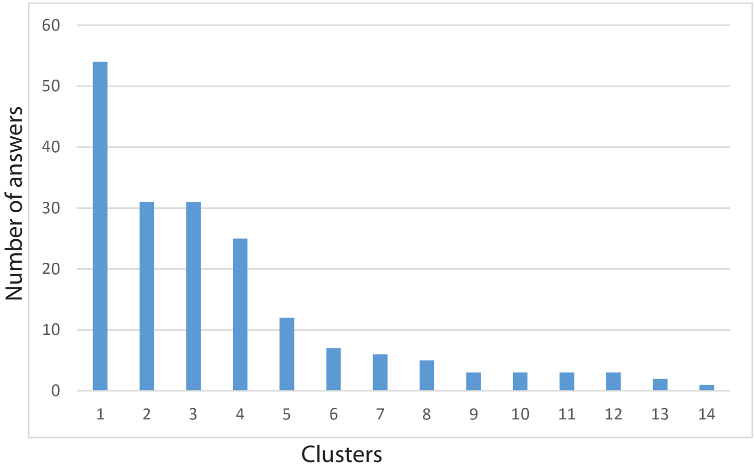


Fig. 5. The distribution of the number of answers in particular clusters

What refers to the basic question, which is the membership of the same questions to the same clusters, we found that indeed, a huge number of questions belong to the same clusters. It is no surprise, that the key factor that influences that, is the cosine distance between the answers, for different sets. The closer the answers between the groups of students are, the larger chance, that they belong to the same cluster. Uncorrelated answers are unlikely to belong to the same clusters. Thus providing conditions, in which the collected data are highly correlated for subsequent years, is the key factor to guarantee high reliability of the data model. The number of answers, that matches particular clusters, of course depends of the free parameter - the cut point in the hierarchic clustering. The lower the cut point, the more detailed clustering, and the more mismatching answers. Together with increasing the cut point, the number of matches rises.

7 Conclusions

The paper investigated the problem of data integration on the example of data coming from student surveys. For this purpose we defined a semantic space, which allows for computing the similarity between the survey answers. This concept allows for identifying answers with close meaning, which is the first step to integrating the data.

The main focus of this paper was to verify if this approach is reliable, and could be used for integrating this kind of data. The correlations obtained for two subsequent years, for which the survey was conducted, indicate that indeed - the statistical distributions for particular questions exhibit high similarities. Thus the approach can be the basis for data integration. Although there are still some question/answer pairs, which do not correlate well. We indicated the ways of dealing with the situation to improve the results.

The other open question is the clustering method to be used to group the answers. In this paper we used the hierarchic clustering, because of its flexibility in steering the granularity with the cut point of the clustering hierarchy. But also other methods should be tested. This is especially important when we realize, that the radius of the clusters could be an important factor influencing the size of particular clusters. Unfortunately in hierarchic clustering we have no direct influence on the radius.

The presented methodology of data integration was demonstrated on a particular application example, but its nature is universal. It can be applied to any kind of data, where we have information about a group of entities, and the entities can be classified into a number of categories. This is a very wide category of problems, so there is a lot of work to do, to analyze the results delivered by our methodology.

Acknowledgments. This work is the result of a project titled “Intelligent methods of analyzing chances and threats in the educational process” (project no. UDA-RPPK.01.03.00-18-019/12-00), co-financed by the European Union from the European Regional Development Fund and from the Budget within the Regional Operational Programme for the Podkarpackie Region for the years 2007-2013.

References

1. Bellahsene, Z., Bonifati, A., Rahm, E.: Schema matching and mapping. Springer-Verlag, Berlin Heidelberg (2011)
2. Euzenat, J., Shvaiko, P.: Ontology matching. Springer-Verlag, New York (2007)
3. Halevy, A., Rajaraman, A., Ordille, J.: Dataintegration: The teenage years. In: 32nd International Conference on Very large Databases, pp. 9-16. VLDB Endowment, Seoul, Korea (2006)
4. Jaszuk, M., Mroczek, T., Fryc, B.: Identifying Semantic Classes within Student’s Data Using Clustering Technique. In: 3-rd International Conference on Data Management Technologies and Applications DATA 2014, pp. 371–376, SCITEPRESS, Vienna (2014)

5. Kaladevi, R., Mrinalinee, T.T.: Heterogeneous Information Management Using Ontology Mapping. ARPN Journal of Engineering and Applied Sciences, 10(5), 2078–2081 (2015)
6. Mao, M.: Ontology Mapping: Towards Semantic Interoperability in Distributed and Heterge-neous Environments. Ph.D. dissertion, Pittsburgh Univ., Pittsburgh, PA. (2008)
7. Murtagh, F., Legendre, P.: Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion?. Journal of Classification, 31(3), 274–295 (2014)
8. The R Project for Statistical Computing, `http://www.r-project.org/`
9. Shvaiko, P., Euzenat, J.: Ontology matching: state of the art and future challenges. IEEE Trans. on Knowledge and Data Engineering, 25(1), 158–176 (2013)