Predicting drop-out from social behaviour of students

Jaroslav Bayer, Hana Bydžovská,
Jan Géryk, Tomáš Obšívač
Computer Systems Unit
Faculty of Informatics, Masaryk University
Brno, Czech Republic
{bayer, bydzovska, geryk,
obsivac}@fi.muni.cz

Lubomír Popelínský
Knowledge Discovery Group
Faculty of Informatics, Masaryk University
Brno, Czech Republic
popel@fi.muni.cz

ABSTRACT

This paper focuses on predicting drop-outs and school failures when student data has been enriched with data derived from students social behaviour. These data describe social dependencies gathered from e-mail and discussion board conversations, among other sources. We describe an extraction of new features from both student data and behaviour data represented by a social graph which we construct. Then we introduce a novel method for learning a classifier for student failure prediction that employs cost-sensitive learning to lower the number of incorrectly classified unsuccessful students. We show that the use of social behaviour data results in significant increase of the prediction accuracy.

1. INTRODUCTION

One of the current trends in higher education is the substantial increase of the first-year students and, consequently, the volume of educational data. Thousands of students are admitted to study at universities every year. They reach interim results, pass or fail at exams, communicate with each other during their studies and many of them fail to finish their study successfully. University staff would like to encourage such students to finish their studies but it is hard to identify them early also because of the huge number of enrolled students. It is important to explore methods that can extract reliable and comprehensive knowledge from the student data that allow prediction of a drop-out with a sufficiently high accuracy.

In this work we utilized student data that have been stored in the Information System of Masaryk University (IS MU), which stores educational data and comprises of all information about students and their studies, about teachers and courses, and also provides examination management tools, excuses registration system, evaluation of on-line tests, and various forms of communication, e.g. discussion boards. We utilized only a subset of information stored in IS MU that is relevant for prediction of the student success, like capacity-to-study test scores, gained credits, average grades, or gender. Data from IS MU are periodically imported to data warehouse Excalibur [3] that combines three main disciplines of data processing—data management, data mining (DM), and visual analytics.

IS MU also stores the complete history of users' requests to the system. Data about students' social behaviour, such as intensity of interpersonal communication or number of mutually shared files, can be observed and stored either immediately, when the particular system function is used, or later from the complete history of users' requests that is present in the form of the system access log. Relations among students (identified from their social behaviour) are main building blocks of a latent social network. With the help of Social Network Analysis (SNA) [4] we compute several new features of a student from the network, for example neighbours characteristics.

In this paper we introduce a novel method for data generation, pre-processing, and educational data mining (EDM) [1; 14; 10] that utilize both the student records and the data about their social behaviour. We show how to predict student drop-out and school failure using DM [7] methods and SNA. We use SNA for creating new study-related features that can help conventional learning methods to increase the accuracy of predicting student performance or detecting a possible drop-out. We intend to build classifiers for early detection and long term prediction of a potential drop-out. The early detection implies a need for the history of data.

Preliminary results for this task were published in doctoral workshop [2]. The highest measured accuracy was above 80% when only student data were employed. We enriched the student data with the data about social behaviour and achieved an increase of the overall accuracy of about 10%. In both cases, the information gain based machine learning (ML) methods generated the most successful classifiers.

Another approach to the prediction of a student study performance that is based on questionnaires can be found in [12]. In [15], a design of a web based system for solving issues related to student performance in higher education is proposed. It utilizes a quality function deployment in combination with DM methods. A novel ML method predicting drop-out in distance higher education from imbalanced datasets is discussed in [9]. It reveals limitations of the existing methods and proposes another approach based on local-cost sensitive techniques. A novel approach to identify factors influencing the student success is discussed in [11]. It focuses on factors available before the beginning of a students degree program suggesting associative rules for subgroup discovery to predict possible drop-outs. A significant improvement of prediction of freshmen drop-out using cost sensitive learning is described in [5]. The highest accuracy of classification was achieved using decision trees. In comparison with our approach of utilizing social behaviour, a combination of data mining methods with natural lan-

guage processing, especially text mining, was employed in [17] to increase the student retention.
In the following section, we introduce the structure of both the student data and the social behaviour data and the necessary preprocessing steps. We describe how we built the social network and applied the analytical methods in Section 2.2. Section 2.3 describes the DM method used for drop-out prediction. In Section 3 we demonstrate the results and the improvement of the classification by measuring the amount of the additional data explored by SNA. Then we show that high-accuracy classifiers can be created for every student regardless of the actual stage of the study. Discussion of results is in Section 4. Finally, we conclude this paper with an overview of the main results and future work in Section 5.

2. DATA AND DROP-OUT PREDICTION

2.1 Student data

Our research considers bachelor students of Applied Informatics admitted to Faculty of Informatics, Masaryk University in years 2006, 2007, and 2008. For that period we can obtain data that match the whole length of the standard bachelor study, i.e. three years. The year 2006 as the lower bound is set as the year when social behaviour data have been started to collect. We explored only the students that were in contact with the school community. Such students produce social behaviour data characterizing them in the university setting.
We selected only general attributes of studies to be able to apply our approach to students of any faculty. To predict a drop-out through the whole period of the study we collected data snapshots for each term of student studies. The set of attributes can be divided into three categories according to the type: Student-related attributes, Semester-related attributes and Attributes related to other studies.

Student-related attributes comprise of the following:

(1) **gender**

(2) **year of birth**

(3) **year of admission**

(4) **exemption from entrance exam**

(5) **capacity-to-study test score**—a result of the entrance examination expressed as the percentage of the score measuring learning potential

Semester-related attributes are the following:

(6) **the number of finished semesters**

(7) **recognized courses**—the number of related courses finished in other studies

(8) **recognized credits**—the number of credits gained from recognized courses

(9) **credits to gain**—the number of credits to gain for enrolled but not yet finished courses

(10) **gained credits**—the number of credits gained from finished courses

(11) **uncompleted courses**—the number of courses a student has failed to complete

(12) **second resits done**—the number of the utilized second resits. Each student can exercise the right to the second resit for only as many times as the standard length of the study in years increased by one.

(13) **excused days**—the number of days when a student is exused

(14) **average grades**—the average grade computed from all gained grades

(15) **weighted average grades**—average grades weighted by the number of credits gained for courses

(16) **the ratio of the number of gained credits to the number of credits to gain**

(17) **the difference of gained credits and credits to gain**

Because a student can be enrolled in more studies or also on more faculties, we added also attributes related to other studies of the student. This set of attributes consists of the following:

(18) **the number of parallel studies at the faculty**

(19) **the number of parallel studies at the university**

(20) **the number of all studies at the faculty**

(21) **the number of all studies at the university**

Data that consist of values of all attributes characterizing a study in a point of time have been extracted from Excalibur. The data set contained 775 students, 837 studies and 4,373 examples in total—one example per a term, where the number of terms for a student varied from 1 up to 8.

2.2 Social behaviour data

The aforementioned set of 775 students is the core of ego-centered social network. We create it from the students plus their direct schoolmates and relations among them. Relations reflects the patterns of social behaviour data. Then we compute new student features from the network structural characteristics and student direct neighbours attributes.
To obtain knowledge concerning a student from perspective of his or her engagement in the school community, we construct a sociogram, a diagram which maps the structure of interpersonal relations. Such social graph allows to find new features by link-based ranking.
There are number of interpersonal ties already evaluated to enhance IS MU full text search. We compute them either online or through system log processing and store them both in the search engine index as a relevant document non-textual tokens and as a part of the user model. These are then used to better order the search results by matching documents (e.g. e-mails, files, courses) related to the respective users [16].

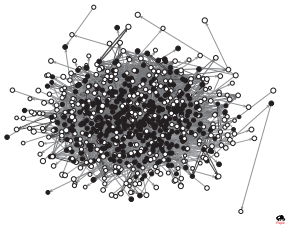

Figure 1: Network with vertices arranged by Kamada-Kawai energy layout algorithm. Dark nodes represent students with successfully finished studies.

Some ties are intuitive and strong facts, namely:

(a) *explicitly expressed friendship*

(b) *mutual e-mail conversation*

(c) *publication co-authoring*

(d) *direct comment on another person*

Weaker ties are more hidden and are derived from the following facts:

(e) *discussion forum message marked as important*

(f) *whole thread in discussion forum or blog marked as favourite*

(g) *files uploaded into someone else's depository*

(h) *assessments of noticeboard's messages*

(i) *visited personal pages*

We measure the value of a tie by its importance and weight it by the number of occurrences. For instance, a tie representing exchange of several e-mails have grater value than a visit to somebody's personal profile. The identification of the best weights is a possible subject of future evaluation. Another notable property of a relation is its direction. It indicates the source and the target of an action which we count as the relation. For example, a person who sent/received an e-mail or who uploaded/received a file in the source/target respectively. Some actions have no direction, e.g. marking the same discussion thread as a favourite one.
As a result we calculated a single number from all mentioned ties reflecting the overall strength of a student relation to any given schoolmate. We found 13,286 such connections representing graph arcs (oriented lines) valued by this number.
Now, the network can be visualized for exploratory analysis of its properties. For example, after applying Kamada-Kawai energy layout algorithm [8] (Fig. 1), we can see that the successful students (black nodes) occupy the area in the middle of the network and are rarely seen on the periphery. In opposite, the unsuccessful ones (white nodes) are placed all over the graph. This, along with results presented later, supports our assumption that higher number and stronger ties have a positive impact on the success of the study while absence of the ties predicts a potential to failure.

2.2.1 New feature generation

This single mode social network of students and their interpersonal ties (i.e. homogeneous information network) allows us to explore it not only visually but also by tools for social network analysis, e.g. Pajek [13]. Moreover, previously unseen features of each student may be computed with such tools. The following two types of features are interesting and give us a new insight into the data.
First, features obtained from the network structure are computed from basic structural characteristics, namely the vertex degrees, the summary of incident line values, and the betweenness centrality:

(22) **degree**—the number of lines that incident with a vertex, represents how many relations the student is involved in

(23) **indegree (or popularity)**—the number of arcs coming to the node, it represents for how many other members of the network the student is a subject of interest

(24) **outdegree**—the number of arcs with opposite direction represents an interest initiated by the given student

(25) **sum of incident line values**—to measure also the strength of the ties

(26) **betweenness centrality**—the number of shortest paths from all vertices to all others that pass through given vertex represents student's importance (global to the network)

Second, features obtained from the neighbourhood properties are also important to examine, and we must measure not only the quantity of person's ties but also their quality. In other words, the academic performance of the surrounding students is important, because it would be hard to get advantage from communication with unsuccessful students. We selected four student features from the data set, preferred by their information gain, to calculate averages of the neighbourhood values (ANV):

(27) **capacity-to-study test score ANV**

(28) **grade average ANV**

(29) **proportion of enrolled and fulfilled credits ANV**

(30) **credits per semester ANV**[1]

2.3 Process of drop-out prediction

We aimed at developing an accurate method for drop-out prediction that would also allow predicting the drop-out in an early stage of the study. The method should have minimum of false negatives, i.e. students that have not been recognized to be in danger of dropping-out.
When all the attributes were used the accuracy was poor. That is why we utilized feature selection methods to reduce the dimensionality of the student data extracted from Excalibur data warehouse. We improved the pre-processing method described in [12] by computing the average rank of attributes while eliminating the extreme values.

[1]Surprisingly, when we tried to use these features with weighting using the strength of the corresponding connection, it has not improved the performance of the classifiers.

The goal was to preserve reliability of attributes for classification after the reduction. Therefore we utilized a combination of feature selection/estimation algorithms based on different approaches. We employed three algorithms based on entropy (InfoGainAttributeEval, GainRatioAttributeEval and SymmetricalUncertAttributeEval), an algorithm selecting the minimum-error attribute for prediction (OneRAttributeEval), an algorithm utilizing χ^2-distribution (ChiSquaredAttributeEval), an algorithm preferring attributes highly correlated with the class but with low intercorrelation to others (CfsSubsetEval), an algorithm looking for the smallest subset of attributes having the consistency equal to that of all attributes (ConsistencySubsetEval), and an algorithm assessing attributes by finding the nearest neighbours for a randomly chosen example from every class. It compares the accumulated differences of values of the corresponding features (ReliefFAttributeEval), and we utilized also two filters (FilteredAttributeEval, FilteredSubsetEval). Then, we computed a list of attributes ordered by the average ranks gained from the ordered lists produced by the feature selection algorithms evaluating the significance of the attributes. For every attribute, we skipped the extreme values—the best and the worst evaluations. We reduced the set of attributes to the 22 most relevant and learned the classifiers again. Except for the Naive Bayes (NB) method, all used machine learning methods achieved a higher accuracy. Examples of the removed attributes are the following: being a seminar tutor, the number of password changes, or the number of enrolled courses.

The list of the refined set of attributes in relevance order can be found in Table 1.

Then we computed significant structural characteristics of the social network to gain additional attributes implying social relations among the students.

We employed machine learning methods from Weka on the student data and then on the data that contained also the social behaviour data. To cover all types of machine learning algorithms, we employed J48 decision tree learner, IB1 lazy learner, PART rule learner, SMO support vector machines, and NB classifier. We also employed ensemble learning methods, namely bagging and voting. We utilized cost-sensitive learning (CSM) and then bagging with cost matrix. All methods have been used with default parameter settings. Performance was measured in terms of accuracy (the number of correctly classified examples over the number of all examples) and True Positive Rate (the number of correctly classified examples from the class of unsuccessful students). We used 10-fold cross-validation.

3. RESULTS

First we created a classifier using only the social behaviour data but the accuracy did not raise above 69%, in fact, it was lower than for learning from student data. However, if we added the attributes that described the social behaviour to the student data, we observed an increase of accuracy that reached 11%. Main results can be found in Table 2. In the first column represents the results obtained from Excalibur data warehouse, followed by the results for the Excalibur data enriched by the social behaviour data. The baseline was 58.86%. The highest accuracy was obtained with PART, 93.67%, and the True Positive (TP) rate 92.30%. Accuracy for the data without information about student's social behaviour did not overcome 90% and the best result was obtained with decision tree learner, 82.53%, and the TP rate 78.50%.

The most significant attributes include the ratio of the number of gained credits to the number of credits to gain, and the average of this ratio measured for neighbours weighted by the strength of their relation in the social network. The seven most relevant attributes are presented in Table 1.

Table 1: Seven the most relevant attributes

Order	Avg. Ord.	Attribute
1	1.000	(16)
2	2.000	(14)
3	2.625	(15)
4	4.500	(5)
5	5.625	(17)
6	6.000	(8)
7	7.750	(10)

Table 2: Learning from student data (Excalibur) and student data enriched with social behaviour attributes (With SNA) [%]

	Excalibur		With SNA	
Method	Accur.	TP	Accur.	TP
ZeroR	58.86	–	58.86	–
NB	77.57	73.5	72.26	83.4
SMO	79.17	64.6	81.59	74.2
IB1	78.14	72.5	89.80	86.2
PART	82.44	73.7	93.67	92.3
OneR	75.89	57.9	88.45	83.8
J48	82.53	78.5	89.89	88.8

We consider social behaviour data to be a characteristic of a student. Therefore, we learned classifiers only from the social behaviour data without snapshots of student studies data. The baseline was slightly lower than for the student data or the enriched data. The most successful classifier was PART with the accuracy 68.82% and the TP with the rate 70.50%. The results are in Table 3.

Table 3: Learning from social behaviour attributes only [%]

Method	Accur.	TP
ZeroR	50.18	–
NB	64.04	80.6
SMO	63.68	83.5
IB1	60.10	63.5
PART	68.82	70.5
OneR	59.50	57.3
J48	68.34	65.0

Then we analyzed how successful a prediction of a drop-out would be for different time periods. We learned classifiers on interim study results enriched by social behaviour data to recognize drop-outs as soon as possible. Results in terms of accuracy (%) are in Table 4.

Table 4: Learning from student data enriched with social behaviour attributes per semester [%]

	1		2		3		4		5		6		7+	
Method	Accur.	TP	Accur.	TP	Accur.	TP	Accur.	TP	Accur.	TP	Accur.	TP	Accur.	TP
ZeroR	50.18	–	50.25	–	53.87	–	58.56	–	64.02	–	72.20	–	76.77	–
NB	71.45	69.1	78.87	75.8	78.98	80.7	78.77	81.8	78.66	80.2	77.56	76.3	68.60	68.0
SMO	72.40	73.9	81.33	80.2	81.02	77.5	83.22	78.1	83.74	72.3	87.56	67.5	85.48	52.3
IB1	66.48	62.4	70.64	67.2	66.72	61.1	71.40	63.2	74.59	61.0	77.07	53.5	90.93	75.8
OneR	62.84	65.7	77.89	77.3	79.71	74.4	83.56	74.4	81.50	66.7	83.90	60.5	80.58	37.5
PART	70.13	69.5	74.82	74.3	76.20	72.8	76.20	73.1	77.24	69.5	79.51	64.0	91.11	83.6
J48	70.73	71.2	74.82	72.8	75.77	72.5	77.91	72.7	77.64	67.8	80.00	63.2	87.11	68.8

Subsequently, we focused on prediction of drop-outs when the history of data about student studies is employed. All data snapshots were used. Results in terms of accuracy (%) are in Fig. 2. On X axis there is a period of study in semesters (e.g. 3 means that only the data from the first 3 semesters have been used for building the classifier). More details are in Table 5 and Table 6.

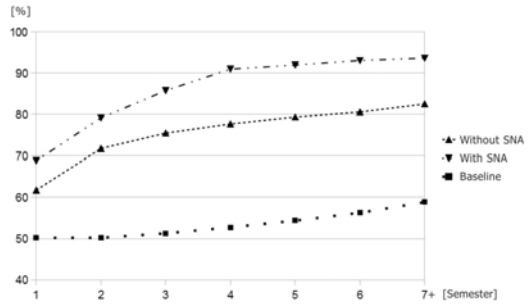

Figure 2: Classifications according to semesters

We can see that for all periods the classification that used only the student data achieves lower accuracy in comparison to the classification on the enriched data. Moreover, starting with the period of the first four semesters the accuracy of classification on the enriched data was higher than 90%. We can conclude that four semesters is a period when our model can predict a drop-out with high probability. We consider this result to be satisfactory. The Masaryk University evaluates the learning potential of students before they are admitted to study.

For our task it is more serious when a student is not recognized to be in danger of a drop-out than the opposite situation. To decrease the number of incorrectly classified unsuccessful students, we tested cost-sensitive learning (CSM) and also bagging, and then bagging with cost matrix, always with the most accurate learning algorithm as the base classifier. In the case of cost-sensitive learning, we set a cost matrix to [0, 1, 0.5, 0] so that the cost of false negative error (i.e. of non-recognized weak students) was twice as high. All the results are in Table 7 in the form of Accuracy (%), TP rate (%), and Incorrectly classified unsuccessful studies (ICUS).

Table 7: Meta-classifiers accuracies

	Accur.	TP	ICUS
Excalibur (J48) CSM	80.45	85.7	258
With SNA (PART) CSM	92.89	92.8	129
Excalibur (J48) Bagging	83.30	87.8	219
With SNA (PART) Bagging	96.66	96.0	55

4. DISCUSSION

Based on the results, we conclude that a student performance appears to be correlated with the social habits, mainly with the frequency of communication. It supports the hypothesis that students with average results but communicating with students having good grades can successfully graduate with a higher probability than students with similar performance but not communicating with successful students. We identified wrongly classified instances and supplemented them with additional information about specific courses. We found that about one third of students did not complete two particular courses (Automata and Grammars and Specialist English). These findings could be useful in the future work.

Classifiers based on the information gain were the most successful ones. The NB classifier suffered from the strong independence assumption, on our data.

We also combined the two most successful classifiers—J48 and PART—and built a meta-classifier where the prediction was computed as the average of probabilities of particular classifiers. However, the overall accuracy was not higher than that of the best classifier.

We investigated the influence of social behaviour data on the accuracy of classification with respect to the gender of students. The additional data did not increase the accuracy at all. Any classifier did not overcome the baseline 92.11%.

In comparison to [12], we employed social network analysis. They achieved higher accuracy but with more specific attributes obtained from the data that was collected specially for the study. These attributes can not be retrieved from standard school information systems, e.g. smoking habits, the parents' level of education, or the number of siblings.

We investigated the influence of cost sensitive learning on the accuracy of a drop-out prediction. Employing a cost-matrix did not decreased the overall accuracy but slightly improved the TP rate. Using bagging with a cost matrix increased both the accuracy and the TP rate. In the case of classification on the student data, the accuracy remained almost unchanged, but the TP rate increased from 78.5% to 87%. The most significant improvement was achieved in the case of classification on the enriched data. The meta-classifier increased the accuracy to 96.66% and the TP rate to 96%. The number of data snapshots of incorrectly classified unsuccessful students decreased from 146 to 55 in the case of the classification using PART. The number of all data snapshots is 4,373.

Table 5: Learning from student data only according to semester [%]

	1		1–2		1–3		1–4		1–5		1–6		All	
Method	Accur.	TP	Accur.	TP	Accur.	TP	Accur.	TP	Accur.	TP	Accur.	TP	Accur.	TP
ZeroR	50.18	–	50.21	–	51.28	–	52.74	–	54.37	–	56.28	–	58.86	–
NB	63.80	34.5	70.56	50.5	72.47	55.0	74.66	59.1	75.82	67.4	76.64	72.7	77.57	73.5
SMO	69.41	64.7	72.62	61.9	75.26	63.1	76.58	64.9	77.64	65.5	78.41	65.4	79.17	64.6
IB1	62.72	61.2	66.38	66.4	69.43	67.0	70.96	68.6	72.30	68.8	74.73	70.2	78.18	72.3
OneR	55.56	41.0	64.93	68.1	70.63	76.5	74.14	79.1	75.32	76.0	75.27	70.9	75.90	57.9
PART	65.35	73.4	71.29	71.5	76.33	71.8	78.97	73.3	80.01	75.0	81.34	77.9	82.44	73.7
J48	61.77	62.8	71.77	73.0	75.47	73.6	77.67	75.2	79.34	75.5	80.61	77.1	82.53	78.5

Table 6: Learning from student data enriched with social behaviour attributes according to semester [%]

	1		1–2		1–3		1–4		1–5		1–6		All	
Method	Accur.	TP	Accur.	TP	Accur.	TP	Accur.	TP	Accur.	TP	Accur.	TP	Accur.	TP
ZeroR	50.18	–	50.21	–	51.28	–	52.74	–	54.37	–	56.28	–	58.86	–
NB	71.45	69.1	75.05	75.4	75.81	78.3	75.41	79.7	75.41	80.7	74.80	80.9	74.07	80.8
SMO	72.40	73.9	77.10	75.7	79.15	76.7	80.10	77.5	80.36	76.4	81.66	76.7	81.68	74.4
IB1	66.43	62.4	67.41	63.7	70.59	67.4	76.92	73.1	81.07	76.8	83.10	79.2	90.10	86.7
OneR	62.84	65.7	69.11	67.0	74.83	74.0	81.27	79.7	83.56	81.5	82.31	79.7	88.20	83.6
PART	70.13	69.5	79.65	77.6	86.60	86.7	90.21	89.3	92.38	90.9	92.99	91.1	93.51	91.9
J48	70.73	71.2	80.01	79.1	84.93	83.0	87.40	85.7	88.77	87.1	88.25	85.8	89.57	87.2

5. CONCLUSIONS AND FUTURE WORK

The main goal of this research was to develop a method for mining educational data in order to learn a classifier to predict the success of a student study and verify the method on real data.

We employed DM and SNA methods to solve the task. We verified the method on students of Faculty of Informatics, Masaryk University but the used data were faculty-independent. Therefore, the method can be used for any unit of a university.

We have shown that structured data gained by means of link-based data analysis increased the accuracy of the classification significantly.

We used only the data that are not specific for a faculty. However, to increase the accuracy of the classification it would be useful to enrich the data with faculty-specific attributes, e.g. information about particular exams that a student passed or failed. Another possible way of future improvement may be to exploit more information from the social network.

Actually, we used only information about a student and his or her direct neighbours. It was intentional because this information is easy to gain and also easy to incorporate into the Information system which is the goal of this research. On the other hand, more complex relations may help further increase the system performance. Data about communication between students and teachers may also be useful. Therefore, we plan to build a heterogeneous [6] network where the vertices will be of more types. Different learning methods can be used then, e.g. multi-label classification.

6. ACKNOWLEDGEMENTS

We thank Michal Brandejs and all colleagues of IS MU development team for the support. This work has been partially supported by Faculty of Informatics, Masaryk University.

7. REFERENCES

[1] R. Baker and K. Yacef. The state of educational data mining in 2009: A review and future visions. *Journal of Educational Data Mining*, 1(1):3–17, 2009.

[2] J. Bayer, H. Bydžovská, J. Géryk, T. Obšívač, and L. Popelínský. Improving the classication of study-related data through social network analysis. In *Proceedings of 7th Doctoral Workshop on Mathematical and Engineering Methods in Computer Science*, pages 3–10. Brno University of Technology, 2011.

[3] J. Bayer, H. Bydžovská, J. Géryk, and L. Popelínský. Excalibur - a tool for data mining. In *Proceedings of the Annual Database Conference - Datakon 2011*, pages 227–228. Brno University of Technology, 2011.

[4] P. Carrington, J. Scott, and S. Wasserman. *Models and methods in social network analysis*. Structural analysis in the social sciences. Cambridge University Press, 2005.

[5] G. W. Dekker, M. Pechenizkiy, and J. M. Vleeshouwers. Predicting students drop out: A case study. In *EDM 2009: Proceedings of the 2nd International Conference On Educational Data Mining. Cordoba, Spain.*, pages 41–50, 2009.

[6] J. Han. Mining heterogeneous information networks by exploring the power of links. In *Proceedings of the 20th international conference on Algorithmic learning theory*, ALT'09, pages 3–3, Berlin, Heidelberg, 2009. Springer-Verlag.

[7] J. Han, M. Kamber, and J. Pei. *Data Mining: Concepts and Techniques*. The Morgan Kaufmann Series in Data Management Systems. Elsevier Science, 2011.

[8] T. Kamada and S. Kawai. An algorithm for drawing general undirected graphs. *Inf. Process. Lett.*, 31:7–15, April 1989.

[9] S. Kotsiantis. Educational data mining: a case study for predicting dropout-prone students. *Int. J. Knowl. Eng. Soft Data Paradigm.*, 1:101–111, January 2009.

[10] A. Kruger, A. Merceron, and B. Wolf. A Data Model to Ease Analysis and Mining of Educational Data. In *EDM2010: Proceedings of the 3rd International Conference on Educational Data Mining. Pittsburgh, USA.*, pages 131–140. www.educationaldatamining.org, 2010.

[11] F. Lemmerich, M. Ifland, and F. Puppe. Identifying influence factors on students success by subgroup discovery. In *EDM2011: Proceedings of the 4th International Conference on Educational Data Mining. Eindhoven, the Netherlands.*, pages 345–346, 2011.

[12] C. Marquez-Vera, C. Romero, and S. Ventura. Predicting school failure using data mining. In M. Pechenizkiy, T. Calders, C. Conati, S. Ventura, C. Romero, and J. C. Stamper, editors, *EDM2011: Proceedings of the 4th International Conference on Educational Data Mining. Eindhoven, the Netherlands.*, pages 271–276. www.educationaldatamining.org, 2011.

[13] W. Nooy, A. Mrvar, and V. Batagelj. *Exploratory Social Network Analysis with Pajek*. Structural Analysis in the Social Sciences. Cambridge University Press, 2011.

[14] C. Romero and S. Ventura. Educational data mining: a review of the state of the art. *Trans. Sys. Man Cyber Part C*, 40:601–618, November 2010.

[15] A. Sahay and K. Mehta. Assisting higher education in assessing, predicting, and managing issues related to student success: A web-based software using data mining and quality function deployment. *Academic and Business Research Institute Conference*, 2010.

[16] M. Čuhel, M. Brandejs, J. Kasprzak, and T. Obšívač. Access rights in enterprise full-text search. In *ICEIS 2010: Proceedings of the 12th International Conference on Enterprise Information Systems, Volume 1: Databases and Information Systems Integration*, pages 32–39. INSTICC, Funchal, Portugal, 2010.

[17] Y. Zhang, S. Oussena, T. Clark, and H. Kim. Use data mining to improve student retention in higher education - a case study. In *ICEIS 2010: Proceedings of the 12th International Conference on Enterprise Information Systems, Volume 1: Databases and Information Systems Integration*, pages 190–197. INSTICC, Funchal, Portugal, 2010.