Use data mining to improve student retention in HE - a case study

USE DATA MINING TO IMPROVE STUDENT RETENTION IN HIGHER EDUCATION – A CASE STUDY

Ying Zhang, Samia Oussena
Thames Valley University, London,UK
ying.zhang@tvu.ac.uk, samina.oussena@tvu.ac.uk

Tony Clark, Hyeonsook Kim
Middlesex University, London,UK
t.n.clark@mdx.ac.uk, hyeonsook.kim@tvu.ac.uk

Keywords: Data Mining, Higher Education, Student Retention, Student Intervention.

Abstract: Data mining combines machine learning, statistics and visualization techniques to discover and extract knowledge. One of the biggest challenges that higher education faces is to improve student retention (National Audition Office, 2007). Student retention has become an indication of academic performance and enrolment management. Our project uses data mining and natural language processing technologies to monitor student, analyze student academic behaviour and provide a basis for efficient intervention strategies. Our aim is to identify potential problems as early as possible and to follow up with intervention options to enhance student retention. In this paper we discuss how data mining can help spot students 'at risk', evaluate the course or module suitability, and tailor the interventions to increase student retention.

1 INTRODUCTION

As the cost of processing power and storage is falling, data storage became easier and cheaper. Universities are facing the immense and quick growth of the volume of educational data (Schönbrunn and Hilbert, 2006). Data mining, sometimes also called Knowledge Discovery in Databases (KDD), can find relationships and patterns that exist but are hidden among the vast amount of educational data. It combines machine learning, statistical and visualization techniques to discover and extract knowledge in such a way that humans can easily comprehend. For universities, the knowledge discovered by data mining techniques would provide a personalized education that satisfies the demands of students and employers.

In order to deliver meaningful analysis, data mining techniques can be applied to provide further knowledge beyond the data explicitly stored. Compared to traditional analytical studies, data mining is forward looking and is oriented to individual students. For example, the clustering aspect of data mining can offer comprehensive characteristics analysis of students, while the predicting function from data mining can help the university to act before a student drops out or to plan for resources based on the knowledge of how many students will transfer or take a particular course.

Student retention is an indicator of academic performance and enrolment management of the university. Poor student retention could reflect badly on the university, and cause serious financial strains. In this paper, we use our project as a case study to discuss how to apply data mining to improve student retention. The rest of the paper is structured as follows. Section 2 introduces related background of the project. An overview about Student Retention is in Section 3. Section 4 discusses the project data source and methodology. Experiment results are discussed in Section 5. Finally Section 6 summarizes this paper.

2 BACKGROUND AND RELATED WORK

Data mining can be applied to a number of different applications, such as data summarization, learning

classification rules, finding associations, analyzing changes, and detecting anomalies (Han et al., 2006, Westphal et al., 1998). Sometimes, data mining has to deal with unstructured or semi-structured data, such as text. Text mining is defined as, "the automatic discovery of previously unknown information by extracting information from text" (Spasic et al., 2005). Data mining is widely applied in many areas such as retail, financial, communication, and marketing organizations.

For universities, data mining techniques could help provide more personalized education, maximize educational system efficiency, and reduce the cost of education processes. It may guide us to increase student's retention rate, increase educational improvement ratio, and increase student's learning outcome.

Gabrilson uses the data mining prediction technique to identify the most effective factor to determine a student's test score, and then adjusting these factors to improve the student's test score performance in the following year (Gabrilson, 2003). Luan uses data mining to group students to determine which student can easily pile up their courses and which take courses for longer period of time (Luan, 2002). These clusters help universities to identify the requirements of each group and make better decisions on how to offer courses and curriculum, required time for teaching and so on. In (Minaei-Bidgoli et al., 2004), authors use data mining classification technique to predict students final grades based on their web-use feature. This can identify students at risk early and allow the tutor to provide appropriate advice in a timely manner.

To understand the factors influencing university student retention, questionnaires are often used to collect data including personal history of the student, implication of student behaviour, perceptions of the student, for example in (Superby et al., 2006) the authors applied different approaches such as decision tree, random forests, neural networks, and linear discriminate analysis to their questionnaires. However, possibly because of the small sample size, the prediction accuracy is not very good. Herzog (2006) collected data from institutional student information system, the American College Test's Student Profile, National Student Clearinghouse, SPSS software are chosen to estimate student retention and degree-completion time. Nearly 50 features including demographics, campus experience, academic experience, and financial aid are applied to predict student retention.
The research shows that decision tree and neural networks performed better when larger data sets are available.

The MCMS (Mining Course Management Systems) project in Thames Valley University (TVU) proposes to build a knowledge management system based on data mining. Different data sources from current university systems (such as library system, student administration, e-learning) are integrated as a data warehouse based on designed data models. Data mining technologies are applied to predict individual student performance as well as the suitability of the course or module. Meanwhile, Text Mining and Natural Language Processing (NLP) technologies are used to generate human friendly result for better understanding (Oussena, 2008).

In MCMS, model-driven data integration is applied to pull data from multiple systems into one data warehouse for reporting and analyzing (Kim, et al., 2009). As data in the data warehouse is cleansed and pre-processed and transformed, it can greatly improve the effectiveness and efficiency of the data mining processes. The knowledge discovered by data mining techniques will enable the university to have more advanced approach in instructing students, predict individual student behaviours and course performance. Text mining and natural language processing technology are applied to some text data sources. They can be used to tailor student intervention. A pilot of this knowledge base system and its intervention strategy for student performance will be running and evaluated for one semester. This will provide an insight on the effectiveness of this approach. Figure 1 shows this general process of the MCMS project.

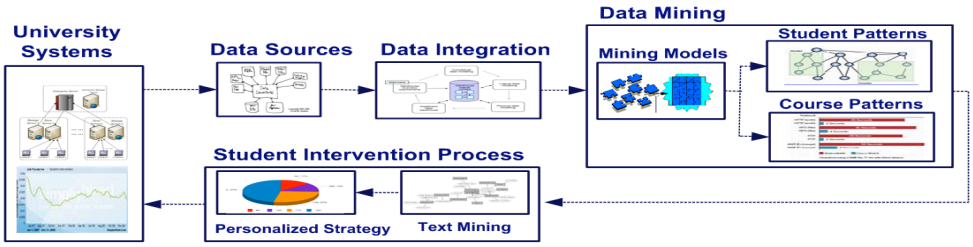

Figure 1: General Process of MCMS.

3 STUDENT RETENTION

One of the biggest challenges that higher education faces is to improve student retention. In general, more students remaining in the university means better academic programs and higher revenue. A report from the UK Parliament Select Committee on Public Accounts (Committee of Public Accounts, 2001) showed that while college participation is about 43 percent, around 28,000 full-time and 87,000 part-time students who started first degree courses in 2004-2005 were no longer in higher education a year later. 91.6 percent of the full-time students entered their second year while only 78.1 per cent were expected to complete (Committee of Public Accounts, 2001). The Higher Education Funding Council for England (HEFCE), the body that distributes government funding to universities in England, links its annual grants to the number of students who stay the university and take, not necessarily pass the exams each year. The sum involved is about £2500 per full time student per year. The government grant loss for UK institutions due to student dropout is about £105m each year (Yorke et al., 2004).

Student fee income also relates closely to student retention. For a medium size university who enrols about 2000 new students each year. If 5% first year students drop out, the fee lost will be increased, if there are international students among them, the fee loss will be much more. Furthermore, dropped out students have a recruitment cost upfront and new students have to be recruited in order to keep university students number steady.

The most widely accepted model in the student retention literature is Tinto's (Tinto, 1995). It examines factors contributing to a student's decision about whether to continue their higher education. It claims that the decision to persist or drop out is quite strongly predicted by their degree of academic integration, and social integration. Tinto argues that from an academic perspective, performance, personal development, academic self-esteem, enjoyment of subjects, identification with academic norms, and one's role as a student all contribute to a student's overall sense of integration into the university (Tinto, 1995). Students who are highly integrated academically are more likely to persist and complete their degrees. The same is true from a social perspective. Student who have more friends at their university, have more personal contact with academics, enjoy being at the university, are likely to make the decision to persist. Poor retention is normally caused by unclear career goals, uncertainty about the course, lack of academic challenge, transition or adjustment problems, limited or unrealistic expectations, lack of engagement, and a low level of integration. According to Tinto, students are most likely to stay on the course if there are close links between their own academic objectives, and the academic and social characteristics of the university. If students find the particular course can combine education and their chosen subject, and greatly help them achieve their goals, their chances of completing would increase dramatically.

There are also other models about student retention. For example, Thomas develops her model "institutional habitus" (Thomas, 2002) based on Tinto's theory, which can be divided by the academic and social experience. The academic experience covers attitudes of staff, teaching and learning and assessment. Different learning styles are supported and diversity of backgrounds is appreciated. Tutors are friendly, helpful and accessible. Assessment gives students the opportunity to succeed, and staffs are available to help. The social experience is about friendship, mutual support and social networks. Thomas noted that one factor in her students' persistence was the fact they felt more at home with their friends.

Seidman developed a formula of student retention (Seidman, 1996) in which:

Retention = Early Identification + (Early + Intensive + Continuous) Intervention

The Seidman formula and shows that early identification of students at risk as well as maintain intensive continuous intervention is the key to increase student retention. He also explains how universities can prepare their programs and courses so students will have the greatest probability of success both personally and academically. It is important to collect family information from students, because this information could aid in a better understanding of individual students. He believes that we could make a difference in helping students attain their academic goals and institutions by increasing their retention rate.

For MCMS, we collect data from different data sources as much as we can to cover Tinto's model, including academic integration, and social integration aspect of students, which are discussed in next section. Seidman's formula can guide us through the whole implement process of MCMS as well. Early identification and early intensive

intervention may make a difference in whether or not the student will leave the institution prematurely.

4 DATA SOURCE AND METHODOLOGY

Thames Valley University (TVU) systems (Oizilbash, 2008) has a great amount of data which can be analyzed and extracted for the data mining system.
Faculties and departments have also important detail data regarding courses and modules which are in document form. This section will discuss each data source and their usage in MCMS.

- Student Record System holds information about student records for example student background, examination results and course enrolment. It is the most important data source for our project.
- Online Learning System allows students to access course material for a particular course module and tutors to extend their classroom teaching using more interactive techniques. This system can help us track the degree of academic integration of students.
- Library System provides information data which could be utilized for student academic integration.
- Reading list system is hosted on the library system server, but has a separated database. It can help us track how often the students borrow books from the recommendation list.
- Online Resource System can be used to identify whether a student is a regular system user.
- Programme Specification is a document which provides course information. Text Mining can be applied to extract course title, learning and teaching method etc.
- Module Study Guide is a text data source that provides module information. It contains details of a module including student assessment strategy, learning outcome and reading lists.
- Course Marketing System is developed for the purpose of marketing usage, which is used to picks a course to be advertised and to research a new course. We can acquire more course information from this system.
- Online Test System enables all to take online entrance skills check. It can help us understand the students' academic background.

MCMS project aims to build a data mining system based on the integration of these TVU systems, which covers the academic perspective, performance aspect of Tinto's model. For example, online learning systems can capture data about students learning behaviour and interaction with the system. The system also includes wiki and group discussion tools which reveal students interaction with their peers and tutors, although little use is made of this data (Oussena, 2008). However we have difficulties to collect other parts of data for Tinto's model, such as student personal development, enjoyment of the subject as well as other social perspectives. We believe the availability of this data will greatly help our research in the future.

Figure 2 shows the system architecture of MCMS. Data sources cover student enrolment, student result, course/module data, learning skills, and student activities etc. Data sources are then integrated and transformed into data warehouse. Data warehouse then generates appropriate data to the data mining engine. As Oracle is the most commonly used database in our data sources, Oracle 11g is chosen as the project platform, which is also integrated with Oracle data warehouse builder as well as Oracle data miner. At this stage, our experiments are mainly based on Oracle 11g.

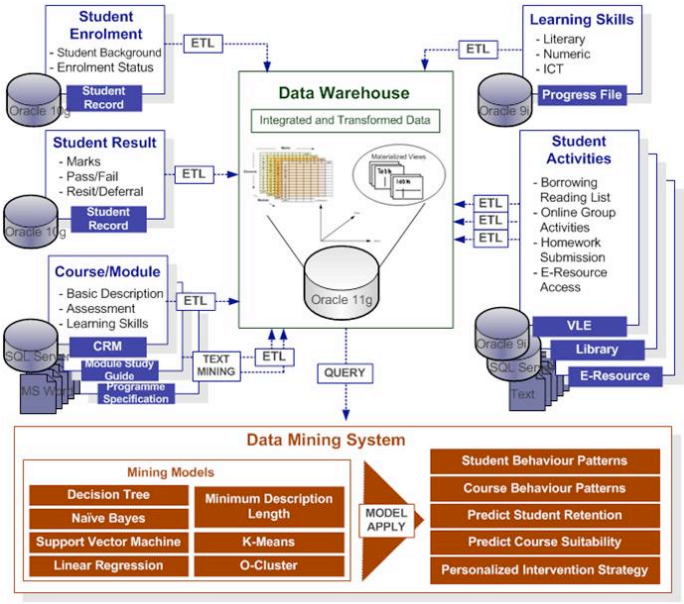

Figure 2: System Architecture of MCMS.

However, in the future we may combine them with other data mining solutions such as weka (Witten et al., 2005) or develop our own data mining models. In MCMS, a model-driven data integration (MDDI) approach is applied in the data integration for our data warehouse. MDDI is a data integration approach that incorporates and proactively utilizes meta data across the data integration process. By coupling data and meta data, MDDI drastically reduces the complexity and provides data integration that is aware of the context of the data. Different modelling approaches have been proposed to overcome every design difficulty of the development of the different parts of a data warehouse system (Mazon et al., 2005 and Fabro et al., 2008).

Once we have the data to identify the characteristics of students who were unsuccessful in past semesters and years, data mining can find the profile pattern of unsuccessful students. For example, feature selection and association rules can help us find the main features that may be related to student dropout. Classification and clustering can identify potential "at risk" students. Text mining and NLP will be used to implement our intervention strategy. This early and intensive intervention can then be measured continuously to see whether or not it has made a difference to the student retention rate. Data mining can also be applied on course data. For example, we could find those modules that are important for a specific course, as they may cause more students to dropout, this will help the university evaluate the module suitability, prepare programs and courses so students will have the greatest probability of success, both personally and academically.

5 EXPERIMENTS

We intend to collect three years historical data from university systems, but currently we only have one year data. For Course marketing system, we have 5,458 records, which include 1,881 courses; 5,352 Course Offering; 7 Schools and 7 Faculties. For Student Record system, there are 4,22,3 Students, 5,352 Course Enrolments. For Library system, there are 144,604 Borrowers, 3,150,816 Loans, 630,190 Items, 435,113 Works and 45,900 Classification. For the Reading List system, there are 552 Course, 1,540 List and 7,084 List Entry. For online Learning systems, there are 2,460 module offering, and 2,021,334 online activities.

For each data source, UML models were developed. A mapping model was also developed that describes how to integrate data from multiple sources. Data Warehousing ETL (Extract, Transform, Load) process is then designed in Oracle Data warehouse builder, and finally the results are inputed to Oracle Data Miner. The data mining process is shown by Figure 3.

Figure 3: Data Mining Process of MCMS.

In order to increase student retention, we should understand why students drop out. Some experiments are used to evaluate Tinto's model. Our student data include: average mark (AVGMARK), online learning systems information (BB_USAGE), library information (LIBRARY_USAGE), nationality (UK), university entry certificate(ENTRYCERTIFICATE), course award (COURSE_AWARD), current study level (CURRENT_STUDYLEVEL), study mode(STUDYMODE), postgraduate or undergraduate (PG_UG), resit number (RESIT_NUM), current year (CURRENT_STUDYYEAR), age(AGEGROUP), gender (SEX), race(RACE) and etc.. Oracle Data Mining provides a feature called Attribute Importance that uses the algorithm Minimum Description Length (MDL) to rank the attributes by significance in determining the target value. As shown by Figure 4 and Figure 5, positive values represents the feature is more significant for dropout than those features with negative values. Thus we can find that whether student drop out is not related to his/ her background, such as age, gender, race etc (as shown in Figure 4), but related to the academic activities, such as how often he/she use online learning system or library system and which year the student is (as shown in Figure 5).

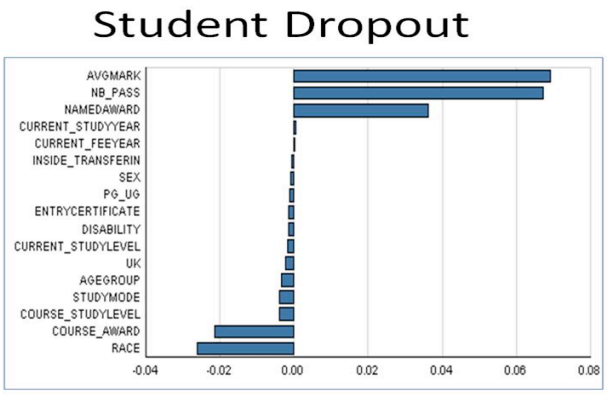

Figure 4: Dropout and student background.

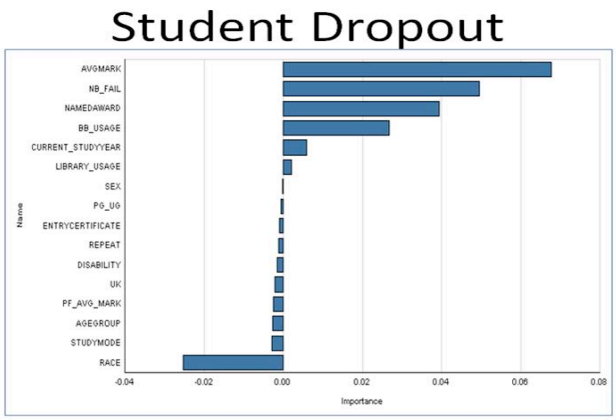

Figure 5: Dropout and student academic activities.

We also identify patterns that describe group of students. There are some interesting finding, such as student transfer to other institution are mostly undergraduate student, enrol with lower certificate, but get higher marks. While student transfer from other institution are mainly international undergraduate student, enrol with higher certificate, use library and online learning system less, and get lower mark. These results can help us understand student behaviour then target efficient intervention strategy to achieve better education results.

An experiment for student dropout prediction is also conducted based on the student profile. The data are divided into training group and evaluation group by the ratio of 2:1. Three algorithms: Naive Bayes (Harry, 2004), Support Vector Machine (Cristianini et al., 2000) and Decision Tree (Quinlan, 1986) are chosen. Different configurations for each algorithm are tested to find the optimum result. As we do not want to give a negative predict error for a real positive target, it is much worse to give a positive error for a real negative target, thus we increase the cost of false negative in the cost matrix. As shown by Table 1, Naive Bayes achieved the highest prediction accuracy while the Decision tree with lowest one.

Table 1: PREDICTION RESULTS.

Accuracy	Naïve Bayes	Support Vector Machine	Decision Tree
Negative	85.9%	78.7%	71.2%

Accuracy			
Positive Accuracy	93.1%	88.3%	91.4%
Average Accuracy	89.5%	83.5%	81.3%

6 DISCUSSION

In this paper, we discussed how to use data mining to improve student retention. For MCMS, the information incorporated into the data warehouse is the historical data of previous students and the features associated with the current and future potential students. We use this information to build the model of student that has potential to drop-out. These students can then be divided into different groups according to their risk value. Once these students are identified, there are several methods which can be taken to improve retention:

Universities should build an intervention programme that will target specific problems. For example, Tinto cites five conditions that best promote retention (Tinto 2000):

- Having high expectations of students.
- Clearly explaining institutional requirements and providing good advice about academic choices. Many students are not clear about their plans, and need help in building a road map.
- Providing academic, social and personal support, particularly in and before the first year.
- Showing students that they are valued. Frequent contact with the staff is important, especially in the first year.
- Active involvement in learning – "students who learn are students who stay". Social learning, where students learn in groups, is particularly valuable, and can help foster friendship, which is another factor that encourages student persistence.

We are currently developing the intervention part of MCMS, which includes a website to monitor student, module, and course information. When a student logs in, he/she can check the detail information of all module results, his/her progress chart, whether he/she is at risk and why. As for tutors, they can login to find detail related module information, which students are at risk and why; if they want to contact the student, they could edit an email which is automatically generated by the system. They can also compare the information of related modules in different semesters. Similarly, program leaders and the head of school can also login to find the information related to their programme.

Different reports with different formats and templates will be sent out automatically by the system as well. For example, students are divided into different group and different intervention rules will be applied to different groups. Students will receive a scheduled email once per month with their detail of their progress of that month in a PDF report. At different times of the semester, emails will also be sent out automatically to provide students with help, such as at the beginning of semester, before exams etc. Triggered emails will be sent out once students' risky values are beyond the threshold, and explain how he/she could make improvement, and who should be contacted. Tutors not only get PDF reports of all their students at scheduled times, but they also receive triggered email once students or modules that may need their attention have been identified.

We are also working to improve the system. For example, we are investigating whether there is a pattern in each individual's reasoning in the decision to withdraw, and how we should implement the intervention programme to increase student retention

REFERENCES

Committee of Public Accounts, 2001-02. *Fifty-eighth Report of Session Improving Student Achievement and Widening Participation in Higher Education in England*, *HC 588*.

Cristianini, N., Shawe-Taylor, J., 2000. *An Introduction to Support Vector Machines and other kernel-based learning methods*. Cambridge University Press.

Gabrilson, S., Fabro, D. D. M., Valduriez, P., 2008. Towards the efficient development of model transformations using model weaving and matching transformations, *Software and Systems Modeling 2003*. Data Mining with CRCT Scores. Office of information technology, Geogia Department of Education.

Han, J. W., Kamber, M., 2006. Data Mining: Concepts and Techniques, 2nd Edition, *The Morgan Kaufmann Series in Data Management Systems*, Gray, J. Series Editor, Morgan Kaufmann Publishers.

Harry, Z., 2004. The Optimality of Naive Bayes, *FLAIRS2004 conference*.

Herzog, S., 2006. Estimating student retention and degree-completion time: Decision trees and neural networks vis-a-vis regression, *New Directions for Institutional Research*, p.17-33.

Kim, H., Zhang, Y., Oussena, S., and Clark, T., 2009. A Case Study on Model Driven Data Integration for Data

Centric Software Development, *In Proceedings of ACM First International Workshop on Data-intensive Software Management and Mining*.

Luan, J., 2002. Data mining and knowledge management in higher education – potential applications. *In Proceedings of AIR Forum*, Toronto, Canada.

Mazon, J. N., Trujillo, J., Serrano, M., Piattini, M., 2005. Applying MDA to the development of data warehouses. *DOLAP 2005*

Minaei-Bidgoli, B., Kortemeyer, G., Punch,W.F., 2004. Enhancing Online Learning Performance: An Application of Data Mining Methods, *In Proceeding of Computers and Advanced Technology in Education*.

Oizilbash. H., 2008. TVU System Overview, *Thames Valley University*.

Oussena, S., 2008. Mining Courses Management Systems, *Thames Valley University*.

Quinlan, J. R., 1986. Induction of Decision Trees. *Machine Learning 1*, 1 pp.81-106.

Schönbrunn, K., Hilbert, A., 2006. Data Mining in Higher Education, Studies in Classification, Data Analysis, and Knowledge Organization Advances in *Data Proceedings of the 30th Annual Conference of the Gesellschaft für Klassifikation e.V.*, Berlin.

Seidman, A., 1996. Spring Retention Revisited: RET = E Id + (E + I + C)Iv. *College and University*, 71(4), 18-20.

National Audition Office, 2007, *Staying the course: the retention of students in higher education*.

Spasic, I., Ananiadou, S., McNaught, J., Kumar, A., 2005. Text mining and ontologies in biomedicine: Making sense of raw text. *Briefings in Bioinformatics* 6(3): 239-251.

Superby, J.F., Vandamme, J-P., Meskens, N., 2006. Determination of factors influencing the achievement of the first-year university students using data mining methods. *Workshop on Educational Data Mining*.

Tinto, V., 1975. Dropout from Higher Education: A TheoreticalSynthesis of Recent Research, *Review of Educational Research* vol.45, pp.89-125.

Tinto, V., 2000. Taking student retention seriously: rethinking the first year of college, *NACADA Journal*, Vol. 19 No. 2, pp. 5-10.

Thomas, L., 2002. Student retention in higher education: the role of institutional habitus , *Journal of Education Policy*, Vol. 17 No. 4, August, pp. 423-442.

Westphal, C., Blaxton, T., 1998. *Data Mining Solutions*, John Wiley.

Witten, I. H., Frank, E., 2005. *Data Mining: Practical machine learning tools and techniques*, 2nd Edition, Morgan Kaufmann, San Francisco.

Yorke, M., Longden, B., 2004. Retention and student success in higher education.*Society for Research in Higher Education*.