Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 3 (2011) 866–871

Procedia Computer Science

www.elsevier.com/locate/procedia

WCIT-2010

# Usage analysis of system for theses acquisition and plagiarism detection

Martin Drlík*[a], Michal Munk[a], Ján Skalka[a]

[a]*Constatntine the Philosopher University, Tr. A.Hlinku 1, Nitra 949 74,Slovakia*

**Abstract**

The national project of The Central Register of the Theses has started in 2008. The project serves as an integrating system for acquisition, archiving and plagiarism detection of the theses from academic information systems of Slovak universities. The first phase of its development has been devoted to the tasks accompanying the processes of acquisition and archiving electronic versions of theses. The national character of the project requires unification of processes associated with theses writing, plagiarism detection and acquisition final versions of the theses from different universities in Slovak republic. The universities, like primary users of this system, have had to adapt their own processes associated with writing, acquisition and archiving of electronic versions of the theses. These inevitable changes have naturally raised many students' and academic staff's questions at universities. The same situation has happened at the Constantine the Philosopher University in Nitra and has led to the development of the helpdesk designed for all stakeholders. The helpdesk has provided relevant and digestedly prepared tutorials and discussion forum about abovementioned changes. The activities of the users have been monitored for the purpose of their further processing and identification of the weaknesses of the theses writing, plagiarism detection and acquisition at the university level. The usage analysis of the presented helpdesk and the segmentation method of its users are discussed in detail in the paper. The segmentation method is based on the monitoring of the users' activities in discussion forums, their searching techniques in available information sources and in posting questions about theses finalization, acquisition and archiving. The authors analyse some aspects of their behavior and discuss interesting findings of the usage analysis of the helpdesk. They give several recommendations for changes in the stakeholder awareness and in the structure of published information materials.



Selection and/or peer-review under responsibility of the Guest Editor.

*Keywords:* usage analysis, user profile, theses acquisition

## 1. Introduction

Thesis writing and commitment is an inevitable condition for successful vindication of each university degree. Limited access to them and some causes of their plagiarism promoted latterly in the media have led to the initialization of the central project of thesis acquisition and originality verification. The main participants of the project are The Ministry of Education of Slovak Republic and the Centre of Scientific and Technical Information. The main actors from the view of the basic processes are The Central Register of Theses (CRT) that integrates data and metadata from universities and provides services for them and Slovak Universities. They are responsible for

1877-0509 
doi:10.1016/j.procs.2010.12.142

theses acquisition and communication between CRT and end-users, i.e. students and advisers. The architecture of the system for theses acquisition and plagiarism detection is described in [1, 2, 3, 4] more details.

Technical part of the processes is relatively simple on the CRT level and has sufficient legislative support, because it contains only few actors and uniquely determined processes. We focus our interests to the processes at the university level because these processes are much more complicated.

One of the greatest problems of theses acquisition comes from the Author's act. According to this law the thesis is considered as a written work created by student for the purpose of acquittal his/her study duties founded in his/her legal relationship with the university. Student can de jure create so defined work with or without license agreement. However, the CRT requires each thesis must be complemented with license agreement by reason of further processing of the thesis.

Entire lifecycle of the thesis can be realized in several, but very much like ways. We can say that Constantine the Philosopher University in Nitra was the first institution in Slovakia where entire lifecycle has been tested. One of the very important advantages of this university is that each step can be realized with the support of information systems of the university:

- Academic information system provides administration associated with thesis announcement and assignment, student's enrolment, thesis review and final exam administration.
- Local Register of Theses collects final theses from the university and communicates with CRT (theses submission to the CRT and accessing to the results of originality verification).
- Information system of Library serves as a final storage for the vindicated theses. It provides access to the theses' content regarding the conditions of the license agreement.
- Helpdesk provides answers for frequent questions and additional sources that support theses writing.

## 2. Helpdesk Support of System for Theses Acquisition

Launching of the project of Central Register Theses and countrywide verification of theses originality give rise to broad discussion among all stakeholders of this system. Each university solves this situation separately because of the differences between them.

Therefore, there was created helpdesk for the support of stakeholders at the Constantine the Philosopher University in Nitra. The helpdesk includes tutorials about thesis writing, templates, and tools for conversion from text formats to pdf. The most important part of the helpdesk is forum that is designed for discussion about miscellaneous facets of abovementioned processes. The students could have asked questions about theses finalization, acquisition and archiving.

## 3. Analysis of Stakeholders' Behavior

The activities of the users have been monitored for the purpose of the further processing and identification of the weaknesses of the theses acquisition at the university level. We used several data mining methods for purpose of analysis of helpdesk stakeholders [5, 6, 7]. We summarize obtained results in the next chapters in more details.

### *3.1. Usage Analysis*

The results are obtained from association rules analysis. This analysis belongs to the non-sequential approaches of data analysis. We do not analyze sequences, but transactions, i.e. we do not include time variable in analysis. The transaction represents a set of problems of one active user (stakeholder) of the helpdesk. We consider categories of problems entered by one user in given time period as one transaction considering obtained data from helpdesk.

We have divided users' problems into these categories:

- *general* – general problems about theses writing, acquisition and archiving,
- *pdf format* – represents category about problems with export final versions of thesis to pdf format,
- *assignment* – problems with creating an electronic assignment,
- *revision* – set of problems about revision of thesis version uploaded previously,
- *out* – set of problems that are not directly related to the thesis writing and acquisition,

- *originality* – important category about problems concerning to verification of originality verification in anti-plagiarism system,
- *license* – collects stakeholders' questions about the intellectual property and authorship
- *structure* – set of answers to questions about the thesis structure.

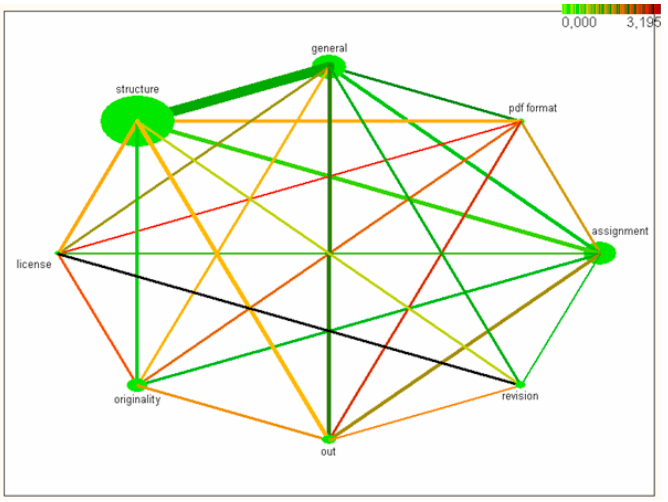


Fig. 1: Web graph – visualization of found rules

The web graph (Fig. 1) visualizes found association rules. The size of node represents *support* of this category item, line width represents support of the rule and line brightness stands for *lift* of the rule. The *lift* parameter - the measure of interestingness - offers the most interesting results because it can be interpreted as how often time categories of problems occur together than in the case if they were statistically insignificant. If *lift* parameter is greater than one then the selected pairs of categories occur more frequently together than apart. It is necessary to say that *lift* parameter does not depend on the rule orientation.

The graph describes digestedly selected associations. We can see that the most frequent categories of problems are *structure* (*support* = 45.5 %) and *general* (*support* = 24.5 %) and their combination (*support* = 14 %). The next interesting finding is that categories *license* and *revision* occur more often together than apart (*lift* = 3.2).

We can apply the same statement for categories *pdf format* and *license*, *pdf format* and *out*, *originality* and *license*, *originality* and *pdf format* (*lift* >2). The *lift* parameter of remaining rules was approximately one.

We deal with sequential rules analysis of obtained dataset. We add time variable to the analysis which means that we make provision for ordering problems of individual users solved in given period. Sequential rules which we obtain from frequent sequences and which satisfy minimal *support* (min *support* = 0.03) can be considered as results of this analysis. The extracted sequential rules summarize Table 1.

Table 1: Table of extracted sequential rules

| Body | ==> | Head | Support (%) | Confidence (%) |
|---|---|---|---|---|
| pdf format | ==> | pdf format | 4.87805 | 66.66667 |
| general | ==> | general | 11.38211 | 48.27586 |
| assignment | ==> | assignment | 8.94309 | 37.93103 |
| structure | ==> | structure | 15.44715 | 35.84906 |
| originality | ==> | originality | 3.25203 | 25.00000 |
| structure | ==> | general | 9.75610 | 22.64151 |
| general | ==> | structure | 4.06504 | 17.24138 |

Notice the rule with the highest *support* (*support* = 15.5 %) and the fourth highest *confidence* (*confidence* = 36.0 %). If the user enters a problem to the category *structure* then she/he enters the problem to the same category with the probability 0.36. Similarly, remaining rules show that next problem of the user belongs to the same category as the first one. The rule *structure ==> general* has the third highest support (*support* = 10 %) and means that if the user enters problem to the category *structure* then she/he enters the problem to the category *general* with the probability 0.23. This probability will be only 0.17 in reverse order.

The rule *pdf format ==> pdf format* represents another interesting finding. If the user enters a problem to this category then he enters the problem to the same category with the probability 0.67. It indicates that if the user has problem with export to format pdf, the problem is serious and user probably has not sufficient ICT skills and need help. Therefore we can recommend preparing additional tutorials for better support of this type of users.

### *3.2. Interaction between categories of problems and weeks*

We created 2-way summary table of observed frequencies absolute observed frequencies/frequencies of interaction as well as relative frequencies. We identified some interesting findings from this crosstabulation that uncover relevant interactions between monitored nominal variables.

We concentrated on frequencies of interaction between categories of problems (*Category*) and weeks (*Week*) at first. The only assumptions for using the test of significance of contingency coefficient are sufficiently great numbers of expected frequencies. This assumption is failed if expected frequencies are less than 5. The validity assumption for chi-square test failed in our case therefore we went not only from the test results but from the graphical visualization of this dependency too. The contingency coefficient is statistically significant and is approximately 0.6 (Table 2). That means that there is a great dependency between the numbers of problems in particular category and observed spaces of time of using helpdesk. The null hypothesis is rejected at the 1 % significance level and so we can say that the number of problems in categories (*Category*) is dependent on the spaces of time (*Week*).

Table 2: Analysis of Crosstabulation - *Category* x *Week*

| | Contingency coefficient | Pearson chi-square | df | p |
|---|---|---|---|---|
| *Category x Week* | 0.5690364 | 228.4151 | 35 | 0.0000 |

Fig 2: Interaction Plot – *Category* x *Week*

Fig. 3: Interaction Plot - *Faculty* x *Category*

Fig. 2 depicts frequencies of interaction digestedly. The interaction plot is represented as categorized polynomial at which each level of category variable is displayed as one polynomial curve. If curves have similar path then

answers are not interdependent and vice-versa. The curves displayed on Fig. 2 have different path so we can say that this graphical representation confirms previous results of analysis.

We observe the highest differences in categories *pdf format*, *originality*, *out* and *revision*. The category *out* represents the set of problems that are not directly related to the purpose of the helpdesk. The waveform confirms the assumption that the stakeholders ask this type of questions during the whole period of using helpdesk about equally often.

### *3.3. Interaction between categories of problems and user's affiliation*

The validity assumption for chi-square test failed in this case too, expected frequencies are less than 5. This implies that we support our statements not only with the value of contingency coefficient (Table 3), but also with the dependency visualization (Fig. 3).

Table 3: Analysis of crosstabulation - Faculty x Category

| | Contingency coefficient | Pearson chi-square | df | p |
|---|---|---|---|---|
| *Faculty x Category* | 0.3544700 | 68.54749 | 35 | 0.0006 |

The contingency coefficient is statistically significant and is approximately 0.35 (Table 3) which means that there is a medium dependency between the stakeholder's affiliation to faculty and the number of given problems in particular category. We can reject the null hypothesis at the 1 % significance level. The number of given problems in particular category (*Category*) depends on stakeholders' affiliation to faculty (*Faculty*). Fig. 3 visualizes this dependency. The curves for 3 faculties (*FSVaZ*, *FSS* and *R*) have various courses. This visualization confirms the results of our previous analysis.

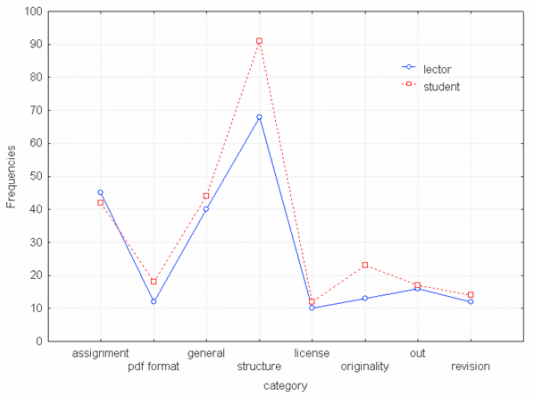


Fig.4: Interaction Plot – *Role* x *Category*

### *3.4. Interaction between categories of problems and roles*

The last examined interaction is between roles and categories of problems. The assumption validation of chi-square test is not failed; expected frequencies are greater than 5. The contingency coefficient (approximately 0.1) is not statistically significant and null hypothesis is not rejected. The number of given problems in particular category (*Category*) does not depend on the role of the helpdesk user (*Role*). This fact confirms the interaction plot on Fig. 4 seeing that the curves have the same course.

## 4. Discussion

The presented results can be interpreted from several points of view. From the helpdesk structure viewing angle is apparent that created helpdesk performs its task and helps users to find relevant answers. The web graph (Fig.1) uncovers the possibility to improve navigation among particular categories of problems.

More consequences emerge from presented analyses for the management of the university and faculties. As we can see on Fig. 3, the concern over the problems of theses writing and their acquisition varies between faculties. It is probably courageously to assume that the stakeholders of two faculties (*FSS* and *FSVaZ*) have not any problems with theses finalization. We suppose sooner that our analysis reveals some problems in information about the existence of helpdesk.

From the thesis advisor point of view it is important to notice that the students deal with the structure of the thesis within two months before the thesis submission. We consider this fact very important because we hold the view that the structure of the thesis should be known earlier. In addition, we can see on Fig. 3 that the problems with structure of the thesis represent the most frequent type of questions in helpdesk. Whence it follows that the advisors should pay more attention to control student during the earlier phases of thesis writing.

We must mention some limitation of our research for completeness' sake. First of all, we analyzed data about active users of the helpdesk. The passive users' behavior has not been analyzed. We can say what particular problem they have seen, but we cannot say if they have read it because we have not monitored the elapsed time on the site.

The second limitation results from the fact that we have analyzed behavior of stakeholders in the first period of real using of the system for theses acquisition and plagiarism detection. Some problems mentioned in helpdesk should be caused by insufficient testing of the system and/or its less user-friendly user interface.

## 5. Conclusion

We described some findings about stakeholders of system for theses acquisition and plagiarism detection and gave several recommendations for changes in the stakeholder awareness and in the structure of published information materials based on the results of usage analysis. We suppose that this analysis contributes positively for the improvement of the process of theses writing and acquisition.

## References

1. J. Skalka, M. Drlík, M. Munk, J. Grman, L.Vozár, The Analysis of Stakeholders' Behaviour on the Helpdesk for Central Register of Theses EDULEARN10, Barcelona : IATED, 2010
2. J. Skalka, Prevencia a odhaľovanie plagiátorstva : zber prác za účelom obmedzenia porušovania autorských práv v kvalifikačných prácach na vysokých školách. Nitra . 2009
3. J. Skalka, M. Drlík, The Plagiarism Prevention and Revelation in Distance Education. Theoretical and Practical Aspects of Distance Learning, Cieszyn, 2009.
4. J. Skalka, M. Drlík, Avoiding Plagiarism in Computer Science E-learning Courses, Problems of Education in the 21st Century : Information & Communication Technology in Natural Scicence Education. 8, (2009).
5. O. Nasraoui, M. Spiliopoulou, J. Srivastava, B. Mobasher, B. Masand (Eds.), Advances in Web Mining and Web Usage Analysis, Springer, 2007.
6. M. Munk and J. Kapusta, P. Švec. Data Preprocessing Evaluation for Web Log Mining: Reconstruction of Activities of a Web Visitor. Procedia Computer Science, 2010, Vol. 1, No 1., pp. 2267-2274. ISSN 1877-0509.
7. Z. Yang, S. Cai, Z. Zhou and N. Zhou, Development and validation of an instrument to measure user perceived service quality of information presenting Web portals. Information & Management, 2005, Volume 42, Issue 4, pp. 575-589. ISSN 0378-7206.