Review report of a final thesis

Czech Technical University in Prague **Faculty of Information Technology**

Student:	Ing. Raúl Carmelo Benítez Netto
Reviewer:	Ing. Tomáš Borovička
Thesis title:	ManaTI: Web Assistance for the threat Analysis supported by Domain Similarity
Branch of the study:	Computer Security

Date: 8. 6. 2017

Evaluation criterion:	*The evaluation scale: 1 to 5.*
1. Difficulty and other comments on the assignment	***1 = extremely challenging assignment,*** *2 = rather difficult assignment,* *3 = assignment of average difficulty,* *4 = easier, but still sufficient assignment,* *5 = insufficient assignment*

Criteria description:
Characterize this final thesis in detail and its relationships to previous or current projects. Comment what is difficult about this thesis (in case of a more difficult thesis, you may overlook some shortcomings that you would not in case of an easy assignment, and on the contrary, with an easy assignment those shortcomings should be evaluated more strictly.)

Comments:
The thesis has more objectives, the first is to develop a software to assist CISCO researchers in analysing web logs and malicious domains. The second objective is to to study WHOIS protocol and research if it is possible to identify similarities among malicious domains based on the WHOIS registration data. Part of the second objective is to develop an algorithm that identifies similarities in WHOIS registration data among multiple domains. This algorithm should be the modular part of the developed software. The last part is to evaluate if the software helps the analyst to process the web-logs more effectively and if the algorithm accurately detects related (similar) domains.
I consider the assignment rather difficult, mainly, because it requires the student to perform a research task as well as develop a fully operational prototype software.

Evaluation criterion:	*The evaluation scale: 1 to 4.*
2. Fulfilment of the assignment	*1 = assignment fulfilled,* *2 = assignment fulfilled with minor objections,* *3 = assignment fulfilled with major objections,* ***4 = assignment not fulfilled***

Criteria description:
Assess whether the thesis meets the assignment statement. In Comments indicate parts of the assignment that have not been fulfilled, completely or partially, or extensions of the thesis beyond the original assignment. If the assignment was not completely fulfilled, try to assess the importance, impact, and possibly also the reason of the insufficiencies.

Comments:
The student developed a software that allows CISCO data analysts to analyse web-logs. This part of the assignment is fulfilled. I consider the other parts of the assignment as not fulfilled due to several fundamental mistakes in the design of the experiments, implementation as well as interpretation of the results.

Evaluation criterion:	*The evaluation scale: 1 to 4.*
3. Size of the main written part	*1 = meets the criteria,* *2 = meets the criteria with minor objections,* *3 = meets the criteria with major objections,* ***4 = does not meet the criteria***

Criteria description:
Evaluate the adequacy of the extent of the final thesis, considering its content and the size of the written part, i.e. that all parts of the thesis are rich on information and the text does not contain unnecessary parts.

Comments:
The thesis has 60 pages. However, content-wise it is not sufficient. The student does not describe the experiments and the software development process sufficiently. From the thesis it is not clear as to how the experiment was performed or how the presented results were achieved.
Description of the software development process is vague. Software development is a process that typically consists of several phases and requires some documentation. In the thesis however, there is no requirements analysis, neither a conceptual design or architecture of the software, there is no mention of a testing of the software having been conducted nor any information about the deployment. Furthermore, there is no user manual for the software.

Evaluation criterion:	*The evaluation scale: 0 to 100 points (grade A to F).*
4. Factual and logical level of the thesis	*0 (F)*

Criteria description:
Assess whether the thesis is correct as to the facts or if there are factual errors and inaccuracies. Evaluate further the logical structure of the thesis, links among the chapters, and the comprehensibility of the text for a reader.

Comments:
The structure of the thesis is not logical, in some parts even confusing; the work lacks a flow, which makes it very difficult to follow. Furthermore, the thesis contains many factual errors, wrong assumption and conclusions. I do not describe all of the mistakes, but rather some examples and overall summary follow.

The introduction is the only chapter that has some structure and it is possible to follow. The second; the most important chapter; unfortunately has a confusing structure and even more confusing content. The author is expected to set goals, explain the conceptual design of the experiments and then describe all the required steps on a theoretical level. When all the theoretical basics are set he should describe the implementation and results achieved. The author here is mixing parts related to implementation with the theory and results (e.g. 2.2.0.1). Terms are often used in sections before they are actually defined. I see as a fundamental problem, that the author does not know the difference between classification and regression and in the thesis he uses regression models for classification tasks. Citation from the thesis: "The classifiers analyzed were Linear Regression and Polynomial Regression...". It is actually unclear what the author wants to achieve since there is no description as to how those models are used. The author is describing the models (mostly copied from the cited source) but does not describe what the inputs are, and neither the outputs of the model in this case! In the results of the experiment (e.g. 2.3.2.4) the author presents a score of the algorithm, however there is only very vague information about the meaning of the score. In the sections 2.3.2 and 2.3.3, one mistake follows the other. It is practically impossible to follow what the author does or even wants to do. In summary, the entire second chapter is confusing and does not describe the work that should be done properly.
Chapter ManaTI Software should describe the development of the software. As already mentioned, software development is a process that typically consists of several phases and requires some documentation. Those parts are missing, the author just describes some functionality in a very unstructured manner.
Moreover, one of the requirements in the assignment is that the software allows external modules. This functionality seems to be supported by the software, however there isn't any description available for an external user. No one can plug-in a module if there is no description of the API and some documentation.
The section, where the effectivity of the analysts with and without the use of the tool is evaluated, is confusing as well. Conditions of the evaluation are not properly set or are rather unclear and the results presented seem to be contradictory. For the group A the average labels per second is 17 and the average labels per minute is 31?

Evaluation criterion:	*The evaluation scale: 0 to 100 points (grade A to F).*
5. Formal level of the thesis	*10 (F)*

Criteria description:
Assess the correctness of formalisms used in the thesis, the typographical and linguistic aspect s, see Dean's Directive No. 14/2015, Article 3.

Comments:
Formal level of the thesis is way below average. There are numerous significant grammatical errors. Inappropriate sentence structures, wrong use of articles, and incorrectly used words are very common throughout the thesis. The student's inability to express precisely has a serious impact on the quality of the work. The sentences often do not make any sense and the reader has to guess what the author means. I present only a few examples (of many mistakes):

Page 4: Reference to a non existing figure.
Page 10: "As far we know, does not exist a software to assist analysts like ManaTI does it. There is much software to visualize weblogs, but any of them provide tools to analyze them."
On page 13 the sentence just ends with four dots: "Detecting malware domains using WHOIS information is possible to some.... Therefore this"
Page 15: There are no axes in the Figure 2.1 (a), which makes it not interpretable in this case.
Page 18: "The results obtained can be appreciated in Figure 2.2."
Page 20: "The used dataset is a comparison between all the domains available for this thesis."
Page 33, 36 Figures overflow.
Page numbers 43, 59 are aligned to the center .
The full description of the figures is in their caption, which makes the size of the captions enormous.
Some figures are not referenced anywhere in the text of the thesis.
There are no references to sources of figures that are not created by the author.

Over all, the formal level of the thesis is not acceptable.

Evaluation criterion:	*The evaluation scale: 0 to 100 points (grade A to F).*
6. Bibliography	*50 (E)*

Criteria description:
Evaluate the student's activity in acquisition and use of studying materials in his thesis. Characterize the choice of the sources. Discuss whether the student used all relevant sources, or whether he tried to solve problems that were already solved. Verify that all elements taken from other sources are properly differentiated from his own results and contributions. Comment if there was a possible violation of the citation ethics and if the bibliographical references are complete and in compliance with citation standards.

Comments:
The author refers to many sources, however, there are only a few technical articles of good quality mentioned. Most of the sources are online. Moreover, the bibliographical references are not complete and not in compliance with citation standards! Due to the incompleteness of the references it is not possible to easily find most of the cited sources.
In the research section of the thesis, there should be a deeper study of existing approaches.

Evaluation criterion:	*The evaluation scale: 0 to 100 points (grade A to F).*

7. Evaluation of results, publication outputs and awards — *50 (E)*

Criteria description:
Comment on the achieved level of major results of the thesis and indicate whether the main results of the thesis extend published state-of-the-art results and/or bring completely new findings. Assess the quality and functionality of hardware or software solutions. Alternatively, evaluate whether the software or source code that was not created by the student himself was used in accordance with the license terms and copyright. Comment on possible publication output or awards related to the thesis.

Comments:
The only valuable results of the thesis may be the software that can be used by CISCO analysts. The question is how it can be extended or further maintained without any documentation. Other results are either incomplete or incorrect.

Evaluation criterion: — *No evaluation scale.*

8. Applicability of the results

Criteria description:
Indicate the potential of using the results of the thesis in practice.

Comments:
No comments.

Evaluation criterion: — *No evaluation scale.*

9. Questions for the defence

Criteria description:
Formulate any question(s) that the student should answer to the committee during the defence (use a bullet list).

Questions:
What is the output of your experimental part?
What is the difference between classification and regression and when are they typically applied?
What was the idea of using regression models in your case?
Why didn't you consider using classification models?

Evaluation criterion: — *The evaluation scale: 0 to 100 points (grade A to F).*

10. The overall evaluation — *25 (F)*

Criteria description:
Summarize the parts of the thesis that had major impact on your evaluation. The overall evaluation **does not** have to be the arithmetic mean or any other formula with the values from the previous evaluation criteria 1 to 9.

Comments:
The assignment of the thesis is difficult mainly because it combines research part and software development part. However, both parts have fundamental mistakes and lack proper description. The thesis does not fulfill the assignment. I hence do not recommend the thesis to pass the defense and I evaluate it with the grade F.

Signature of the reviewer: