2015

The Utility Problem in Case Based Reasoning Alleviated by Big Data

DT228B - ASD

NAOMI SMYTH – C06473075

People Associated – Sarah Jane Delaney

Sources of Data – UCI Machine Learning Repository

1. BACKGROUND, CONTEXT AND SCOPE

Similar problems have similar solutions and the types of problems an entity encounters tend to recur (Leake, 1996).

Case-based reasoning (CBR) is an approach to problem solving and learning in which specific knowledge about previous experiences is used to find solutions to a new, similar problem. In CBR, sustained learning is incremental as new problems are solved, they are added to the existing case-base.

There are four tasks in the CBR cycle; retrieve, reuse, revise and retain. Solving a problem by CBR begins with a problem description. This description is measured against similar problems which are stored in a case base. The similar cases are *retrieved* from the case base and the solutions *reused*. The solution may be *revised* to better adapt to the new problem and finally, this new problem and solution pair is *retained* in the case base CBR learns from experience by retaining the knowledge every time a new problem is encountered.

Modelling human behaviour in cognitive science and developing artificial intelligent systems are the two primary motivations for CBR (Leake, 1996). CBR alleviates a number of issues in AI such as knowledge acquisition, knowledge maintenance, increased problem-solving efficiency, increased quality of solutions and user acceptance.

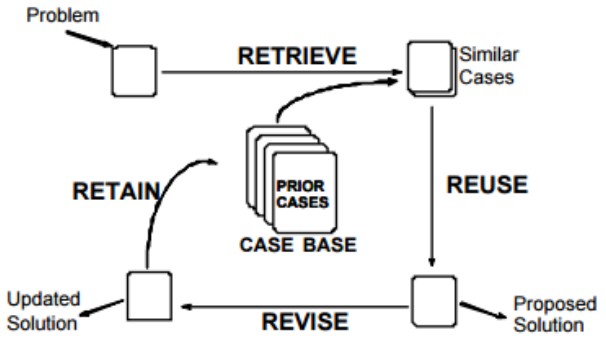


Figure 1: The CBR Cycle (Aamodt and Plaza, 1994)

2. PROBLEM DESCRIPTION

K nearest neighbour (KNN) is one of the most straightforward classifiers used in CBR. The KNN classification method identifies the nearest neighbours to a query and decides the class of the query from these neighbours (Cunningham, Delany, 2007).

In CBR, the addition of new cases to the case base improves the quality of solutions, increases efficiency and allows for greater coverage of problems. However, as the case base grows in size, we encounter an issue known as the *utility problem*. The efficiency of the system will degrade as the *retrieval* task begins to take a very long time. The utility problem is shown to exist when the cost associated with searching for knowledge outweighs the benefit of applying the knowledge (Houeland, Aamodt, 2010).

As the case base grows, the efficiency drops. Once an optimum case base size has been exceeded, there is a trade-off between the quality of the solution and the time required to retrieve the solution. Two factors determine the scope of the problem; the mean retrieval time for a given case-base size and the mean adaptation time for the case-base size. As new cases are added retrieval costs increase and adaptation savings drop (Smyth, 1996).

3. LITERATURE REVIEW

3.1 Literature Review

Francis and Ram proposed to create computational models of the utility problem in case based reasoning in an attempt to identify the root cause and design an effective coping mechanism. The paper concluded that the utility problem occurs on both serial and parallel machines, but was easier to cope with on parallel machines. The coping mechanisms suggested are deletion policies and guided search policies (Francis, Ram, 1993).

A 1994 paper by Aamodt and Plaza gives a comprehensive overview of the foundational issues in case-based reasoning. The methods for case retrieval, reuse, solution testing, and learning are discussed (Aamodt, Plaza, 1994).

A 1996 paper by Leake provides an overview of the process of CBR, the reasons for using CBR and points to new directions to be addressed. The paper suggests that the current generation of CBR systems would cause challenges such as the case adaptation problem, in future research (Leake, 1996).

Smyth and Cunningham state the argument for large case bases in their paper, and provide a thorough

analysis of the utility problem and examine the root causes (Smyth, Cunninghame, 1996).

Munoz-Avila suggests that detrimental retrieval is a more adequate method than adaptable cases in the context of case-based planning. The paper found that Eager Case Retention Policy was too permissive resulting in large case bases and that Retrieval Failures results in decreased competence and increased redundancy. Detrimental Retrieval was shown to be the most effective (Munoz-Avila, 1999).

López De Mántaras et al completed a comprehensive study on case-based reasoning and the problem solving cycle of retrieve, reuse, revise and retain (López De Mántaras, 2005).

3.2 Approaches to Solving Problem

The current approaches to solving the utility problem are either to apply deletion policies or to apply indexing methods.

Smyth and Keane propose a competent deletion policy for case-based reasoning systems to minimise the utility problem. A common machine learning method ensures that the stored knowledge is relevant, and deletes the structures that are not considered useful. The solution uses

"*…a model of case competence to guide the learning and deletion of cases"* (Smyth, Keane, 1995).

Smyth and McKenna propose a new method for constructing compact and competent case bases by allowing cases to be selected on the basis of their individual competence contributions. This method not only applies a deletion policy, but also edits the training data to ensure that the initial case base is near-optimal as all cases in the case base will contribute to performance (Smyth, McKenna, 1999).

Wilson and Martinez provide a review of existing algorithms that are used to reduce storage requirements in instance-based learning algorithms and propose six additional reduction algorithms that can be used to remove instances from the concept description and an analysis of their performance (Wilson, Martinez, 2000).

Case Retrieval Nets are a memory model which apply a spreading activation process to the case base in order to retrieve cases which are sufficiently similar to the posed query case.

Burkhard and Lenz provide a formal description of CRNs and propose it as a suitable method to improve the retrieval step in CBR. They found CRNs supported efficient case retrieval for case bases up to 35,000 cases and CRNs support flexible case retrieval (Burkhard, Lenz, 1996).

In *Case Retrieval Nets Applied to Large Case Bases*, the authors apply CRNs to a large case base and obtain results that suggest that CRNs can successfully handle larger case bases. The case bases used ranged in size from 1,471 to 67,557. The results show that CRNs are able to handle the case bases of smaller sizes, a shortage of memory was observed storing more than 40,000 cases. They found the CRNs required 10 percent less retrieval time than a linear search (Lenz, Burkhard, 1996).

3.3 Gaps in Research

The research into Big Data Platforms alleviating the utility problem has not been sufficiently completed. In the most recent paper from Jalali and Leake, they implement a big-data version of ensembles of adaptation for regression, using MapReduce to illustrate the practicality of this solution. The results were encouraging for the application of big data methods to the complete CBR process. The next direction of this research is to compare accuracy and speed performance of traditional methods and big data methods for CBR (Jalali, Leake, 2015).

4. RESEARCH QUESTION

"Can Hadoop, implementing MapReduce, improve the performance of retrieval in Case Based Reasoning to alleviate the Utility Problem?"

5. HYPOTHESIS

The Utility problem which presents in case-based reasoning for very large case bases can be better addressed using big-data platforms such as Hadoop which implement the programming model MapReduce compared to current indexing techniques such as Case Retrieval Nets. If both methods are applied to the same very large case base which has been shown to exhibit the utility problem, the big-data method will show better performance than the CRN method.

The objective of the research is to show that current solutions for alleviating the utility problem are not effective and the use of a big data platform will show better performance. The utility problem will be shown to exist by first carrying out the k-Nearest Neighbour method on a very large case base. The CRN method will be applied to the same case base and will show some improvement but still exhibit the problem. The big-data method will then be applied and the increase

in performance will be measured by the time it takes to complete the retrieval task.

The research methodologies used are quantitative. The three scenarios will measure the time taken to retrieve cases from the case base for a given query and the results will be plotted on a graph for comparison. The results will clearly show the difference in performance for the three methods.

Secondary research is ongoing to complete a comprehensive literature review of the previous research already completed on case based reasoning, the utility problem and current solutions.

6. DESIGN AND IMPLEMENTATION

The study will be carried out in three parts. A very large case base is required so that results can be collected for many different sized case bases. A paper investigating CRNs found performance issues for case bases of 40,000 cases, a data set of at least 1 million will be used so that the case-base size can be gradually increased to show how the performance degrades. The content of the case base is not that important, the only requirement is that it is large enough.

The first stage will use the k-Nearest Neighbour classifier on the case base to show the existence of the utility problem. The next stage will apply the CRN method to the same case base to show some improvement on the linear approach but will still exhibit the utility problem for larger case-bases. The final stage will apply the big data method to the case base. The big data platform that will be used will be Hadoop, and it will use the programming model MapReduce to execute the CBR retrieval step.

The first step will use a very small sample of the dataset, one which will not suffer from the utility problem. The size of the case base will be increased for each new retrieval step and the performance measured as the time taken to retrieve the cases from the case base. The utility problem will occur as the case base gets larger and performance will deteriorate. The time taken to retrieve cases from the case base will be recorded to measure performance and the degradation in performance can be described with a graph which shows time taken on the x-axis and number of cases on the y-axis.

CRNs will be used on the same sample sizes with the same query so that the performance can be measured on retrieval time. There will be an improvement in performance, as has been shown in previous literature, but as the dataset becomes very large the utility problem will still occur. The data will be plotted on the same graph as the kNN results to show the difference in speeds between the two methods.

MapReduce is a programming model which can process large datasets. A map function is specified which processes key/value pairs to generate a set of intermediate key/value pairs and so would be an ideal solution for a case-based reasoning problem. Hadoop MapReduce can process very large amounts of data in parallel on large clusters. The same query will be carried out using this method on the same sample sizes and the time take to retrieve the cases recorded. The performance of the three methods can then be compared by plotting the results of this stage of the experiment alongside the previous two graphs.

The UCI repository of machine learning has several data sets that are larger than 1 million which would be suitable for the proposed evaluation. One such data set is the Knowledge Discovery and Data Mining Tools Competition Data for 1999. This dataset contains simulated intrusions in a military network environment and the problem to solve would be whether a new connection would be considered “good” or “bad” based on the cases in the database. The data set has 4 million entries, which will be sufficient for the problem proposed.

7. EVALUATION OF DESIGN

The evaluation will be completed on the three stages of the experiment. The performance was measured as the time taken to retrieve the cases from the case base. The improvement in performance can be calculated and a complete statistical analysis will be performed on the results. The content of the dataset used for testing is not that relevant to the study but a description of the content is still required to understand the query that will be carried out on the case base and the cases that are retrieved. Consideration will need to be taken to make sure the accuracy of the results is constant for all stages of the experiment so that performance can be measured by speed alone, if the accuracy of the results varies this will need to be used as a measurement also.

The hypothesis will be accepted if the results show the big data platform performed better than both the CRNs and kNN classifier. It is thought that the big data platform will show a significant improvement on the CRNs method as the literature has shown that this method still suffers from the utility problem when the case base gets into the tens of thousands. If the big data platform does not exhibit the utility problem for the case base of one million entries than the utility problem

can be said to be alleviated. If this solution were a viable option there would be no requirement for deletion methods to reduce the size of the case base.

The findings can be related to the research question as we can show that we have alleviated the utility problem using a big data platform with the results. The results will give us a comparison of how a big data platform performs against CRNs and kNN. A statistical analysis will show what percentage difference there is in performance and this will show how much the method alleviates or does not alleviate the utility problem.

8. BIBLIOGRAPHY

Francis, A., Ram, A. (1993) "Computational Models of the Utility Problem and their Application to a Utility Analysis of Case-Based Reasoning" *Proceedings of the Workshop on Knowledge Compilation and Speedup Learning.*

Aamodt, A., Plaza. E. (1994) "Case-based reasoning: Foundational issues, methodological variations, and system approaches." *AI Communications* Issue 7 Volume 1 pp. 39–59.

Smyth, B., Keane, M. (1995) "Remembering to Forget" *Proceedings of the 14th International Join Conference on Artificial Intelligence* pp. 377-382.

Leake, D., (1996) "CBR in Context: The Present and the Future." *Case-Based Reasoning: Experiences, Lessons, and Future Directions"* pp. 1-30

Smyth, B., Cunningham, P. (1996) "The Utility Problem Analysed: A Case-Based Reasoning Persepctive" *Third European Workshop on Case-Based Reasoning.* Switzerland.

Lenz, M., Burkhard, H.D., Brückner, S., (1996). "Applying case retrieval nets to diagnostic tasks in technical domains". *Advances in Case-Based Reasoning* pp. 219-233.

Lenz, M. (1996) "Case Retrieval Nets Applied to Large Case Bases" *Humboldt University 1996 Conference* pp. 111-118.

Lenz, M. and Burkhard, H.D., (1996). Case retrieval nets: Basic ideas and extensions. *KI-96 Advances in Artificial Intelligence* Volume 1137, pp. 227-239.

Munoz-Avila, H. (1999) "A Case Retention Policy based on Detrimental Retrieval" *Case-Based Reasoning Research and Development* Volume 1650 pp. 276-287.

Smyth, B., McKenna, E. (1999) "Building Compact Competent Case-Bases" *Case-Based Reasoning Research and Development* Volume 1650, pp. 329-342.

Wilson, R., Martinez, T. (2000) "Reduction Techniques for Instanced-Based Learning Algorithms" *Machine Learning* Vol. 38, pp. 257-286.

Smyth, B., Keane, M. (2001) "Hierarchical Case-based Reasoning Integrating Case-based and Decompositional Problem-Solving Techniques for Plant-Control Software Design" *IEEE Transaction on Knowledge and Data Engineering* Volume 13, Issue 5, pp. 793-812.

Ontanon, S., Plaza, E. (2003) "Collaborative Case Retention Strategies for CBR Agents" *Case Based Reasoning Research and Development* Vol. 2689 pp. 392-406.

López De Mántaras, R. et al (2005) "Retrieval, reuse, revision, and retention in case-based reasoning" *The Knowledge Engineering Review"* Vol. 20 Issue 3 pp. 215-240.

Xiong, N., Funk, P. (2006) "Building Similarity Metrics Reflecting Utility in Case-based Reasoning" *Journal of Intelligent and Fuzzy Systems* Volume 17 Issue 4 pp. 407-416.

Bridge, D., Goker, M., McGinty, L., Smyth, B. (2006) "Case-based recommender systems" *The Knowledge Engineering Review* Volume 20 Issue 3 pp. 315-320.

Plaza, E., McGinty, L. (2006) "Distributed case-based reasoning" *The Knowledge Engineering Review* Volume 20 Issue 3 pp. 261-265.

Cunningham, P. and Delany, S.J., (2007) "k-Nearest neighbour classifiers" *Multiple Classifier Systems*, pp.1-17.

Plaza, E. (2008) "Semantics and Experience in the Future Web" *Advances in Case-Based Reasoning* Vol. 5239 pp. 44-58.

Houeland, T., Aamodt, A. (2010) "The Utility Problem for lazy learners – towards a non-eager approach" *Case-Based Reasoning Research and Development* Vol. 6179 pp. 141-155.

Jalali, V., Leake, D. (2013) "Extending Case Adaptation with Automatically-Generated Ensembles of Adaptation Rules" *Case-Based Reasoning Research and Development* Volume 7969 pp 188-202.

Hayes, M., Shah, S. (2013) "Hourglass: A Library for Incremental Processing on Hadoop" *2013 IEEE International Conference on Big Data* pp. 742-745.

Jalali, V., Leake, D. (2013) "A Context-Aware Approach to Selecting Adaptations for Case-Based Reasoning" *Modelling and Using Context* Volume 8175 pp. 101-114.

Jalali, V., Leake, D. (2014) "On Retention of Adaptation Rules" *Case-Based Reasoning Research and Development* Volume 8765 pp. 200-214.

Jalali, V., Leake, D. (2015) "CBR Meets Big Data: A Case Study of Large-Scale Adaptation Rule" *Case-Based Reasoning Research and Development* Volume 9343 pp. 181-196.

9. ACTIVITIES

The dataset has already been found, a dataset of 4 million entities will be used from the machine learning repository.

Secondary research is ongoing, about 30 papers have been read in the area of case-based reasoning, the utility problem and big-data platforms. More research into Hadoop and MapReduce is required. Due to finish mid-February.

The Big Data Platform will be hosted on a cloud server such as Amazon Web Services and approval for this expense will be requested first when term recommences in January.

The kNN classifier will be applied to the dataset first, this is a straightforward classifier but it will take time getting sufficient result on such a large dataset and the classifier will be repeated on varying sized samples of the dataset. Due to finish mid-February.

Whilst the experiments are ongoing the programming model MapReduce needs to be learnt over two weeks.

Some programming will be required to execute the CRNs method which will be carried out next. The experiments and analysis will be completed by the end of March.

The MapReduce program will be written as the experiments with CRNs are being carried out in mid-March over one week.

The Big data method will be applied after the results of CRNs has been collected due to be completed by the end of April.

The results will be evaluated and a statistical analysis will be completed and the report will be written up due to be completed by the end of May.

10. GANTT CHART

Disertation

DIT

Today's Date: 22/12/2015 Tuesday

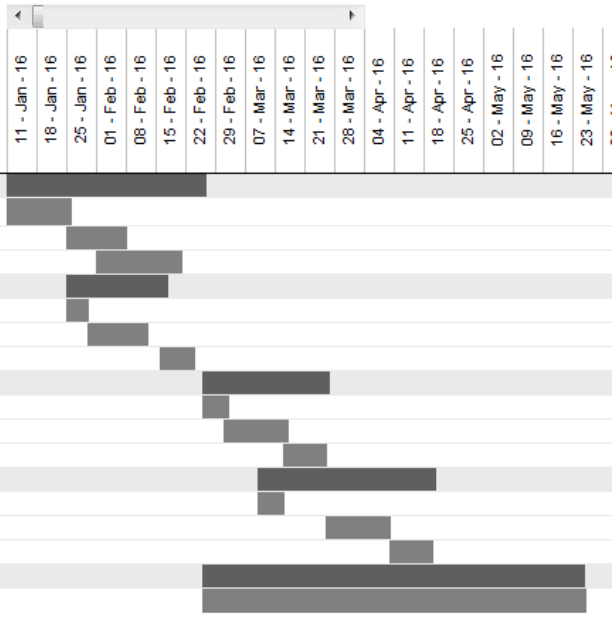
(vertical red line)

Project Lead: Naomi Smyth

Start Date: 10/01/2016 Sunday

First Day of Week (Mon=2): 2

WBS	Tasks	Task Lead	Start	End	Duration (Days)	% Complete	Working Days	Days Complete	Days Remaining
1	**Secondary Research**	Naomi Smyth	11/01/16	26/02/16	47	0%	35	0	47
1.1	Complete secondary research		11/01/16	29/01/16	15	0%	15	0	15
1.2	Learn MapReduce		25/01/16	08/02/16	14	0%	14	0	14
1.3	Write Literature Review		01/02/16	26/02/16	20	0%	20	0	20
2	**kNN Classifier**	Naomi Smyth	25/01/16	25/02/16	24	0%	24	0	24
2.1	Create kNN classifier		25/01/16	29/01/16	5	0%	5	0	5
2.2	Execute experiment		30/01/16	18/02/16	14	0%	14	0	14
2.3	Analyse results		16/02/16	25/02/16	8	0%	8	0	8
3	**CRN method**	Naomi Smyth	26/02/16	26/03/16	30	0%	21	0	30
3.1	Create CRNs		26/02/16	02/03/16	6	0%	6	0	6
3.2	Execute experiment		02/03/16	16/03/16	15	0%	15	0	15
3.3	Analyse results		16/03/16	25/03/16	10	0%	8	0	10
4	**Big-data method**	Naomi Smyth	10/03/16	20/04/16	42	0%	30	0	42
4.1	Create MapReduce program		10/03/16	15/03/16	6	0%	4	0	6
4.2	Execute experiment		26/03/16	09/04/16	15	0%	10	0	15
4.3	Analyse Results		10/04/16	19/04/16	10	0%	7	0	10
5	**Write up**	Naomi Smyth	26/02/16	25/05/16	90	0%	64	0	90
5.1	Write up report		26/02/16	25/05/16	90	0%	64	0	90



11. APPENDICES

Reviewer (Student number) D3123766 Proposer (Student number) C06473075

How can the definition of the domain/scope of the research being proposed be improved? and why?	Domain is topical and interesting.
What should the proposer include in the literature review? and why?	Security publications. Technology.
Are the gaps found in the literature clear and convincing? How can these be clarified/improved?	Yes, slow adaption rates in U.S.A. has left research gaps.
Is the research problem/question clear? How can it be improved?	Rephrase as question. Be more specific.
Can you suggest any relevant research paper/s to the proposer? If yes, provide title/s and author/s.	No.

Reviewer (Student number) C02025027 Proposer (Student number) C06473075.

How can the definition of the domain/scope of the research being proposed be improved? and why?	Domain - Good explanation. What are you trying to achieve? (Scope) Explain theorisation?
What should the proposer include in the literature review? and why?	
Are the gaps found in the literature clear and convincing? How can these be clarified/improved?	Not clear what the gaps are in the existing research.
Is the research problem/question clear? How can it be improved?	Very broad? How will this be tested?
Can you suggest any relevant research paper/s to the proposer? If yes, provide title/s and author/s.	

Reviewer (Student number) D10122492 Proposer (Student number) C06473075

How can the definition of the domain/scope of the research being proposed be improved? and why?	A lot of abbreviations on slide well spoken, Paced well.
What should the proposer include in the literature review? and why?	Good review, think on getting rid of any credit Cards ☺
Are the gaps found in the literature clear and convincing? How can these be clarified/improved?	A lot of detail on this slide to take in.
Is the research problem/question clear? How can it be improved?	
Can you suggest any relevant research paper/s to the proposer? If yes, provide title/s and author/s.	N/A

Reviewer (Student number) D13123578 Proposer (Student number) C06473075

How can the definition of the domain/scope of the research being proposed be improved? and why?	No improvement required
What should the proposer include in the literature review? and why?	Maybe cover some of the alternative security technologies in more detail
Are the gaps found in the literature clear and convincing? How can these be clarified/improved?	Good motivation, maybe link the gaps more clearly to the literature
Is the research problem/question clear? How can it be improved?	Appropriate as spoken, otherwise clear.
Can you suggest any relevant research paper/s to the proposer? If yes, provide title/s and author/s.	Nothing to suggest

Reviewer (Student number) _____ Proposer (Student number) C06473075

How can the definition of the domain/scope of the research being proposed be improved? and why?	
What should the proposer include in the literature review? and why?	
Are the gaps found in the literature clear and convincing? How can these be clarified/improved?	unclear the gaps, because doesn't focuses in only one
Is the research problem/question clear? How can it be improved?	"effectiveness" what kind of effectiveness are you looking for?
Can you suggest any relevant research paper/s to the proposer? If yes, provide title/s and author/s.	

Reviewer (Student number) _____ Proposer (Student number) C06473075

How can the definition of the domain/scope of the research being proposed be improved? and why?	
What should the proposer include in the literature review? and why?	
Are the gaps found in the literature clear and convincing? How can these be clarified/improved?	Other payment methods - iphone touch etc.?
Is the research problem/question clear? How can it be improved?	
Can you suggest any relevant research paper/s to the proposer? If yes, provide title/s and author/s.	Put in a What is & you've got a question define what is "effectiveness"

Reviewer (Student number) D14122975 Proposer (Student number) C0647307S

How can the definition of the domain/scope of the research being proposed be improved? and why?	Clear and well explained.
What should the proposer include in the literature review? and why?	
Are the gaps found in the literature clear and convincing? How can these be clarified/improved?	Convincing, good graphics. ~~may have~~
Is the research problem/question clear? How can it be improved?	A little unclear on exactly what you are going to achieve, how you will define effectiveness ~~and the~~ (I assume less fraud is all not others)
Can you suggest any relevant research paper/s to the proposer? If yes, provide title/s and author/s.	N/A

A good story was painted.
The general presentation was a bit long, but all the details were relivant & interesting.

Reviewer (Student number) D14123716 Proposer (Student number) C0647307S

How can the definition of the domain/scope of the research being proposed be improved? and why?	Domain was clear - explanation of what could not present fraud is needed be good
What should the proposer include in the literature review? and why?	
Are the gaps found in the literature clear and convincing? How can these be clarified/improved?	
Is the research problem/question clear? How can it be improved?	Has tokenisation been implemented - if not how will this be measured
Can you suggest any relevant research paper/s to the proposer? If yes, provide title/s and author/s.	

Reviewer (Student number) C08782610 Proposer (Student number) C06473075

How can the definition of the domain/scope of the research being proposed be improved? and why?	
What should the proposer include in the literature review? and why?	
Are the gaps found in the literature clear and convincing? How can these be clarified/improved?	- You mentioned it would prevent phishing. How would this work? Grant I may have missed how it worked
Is the research problem/question clear? How can it be improved?	Seems like a very good and interesting proposal
Can you suggest any relevant research paper/s to the proposer? If yes, provide title/s and author/s.	

Reviewer (Student number) DL41220 Proposer (Student number) ~~C06473075~~

How can the definition of the domain/scope of the research being proposed be improved? and why?	OK
What should the proposer include in the literature review? and why?	-
Are the gaps found in the literature clear and convincing? How can these be clarified/improved?	-
Is the research problem/question clear? How can it be improved?	- NOT QUESTION NOT CLEAR.
Can you suggest any relevant research paper/s to the proposer? If yes, provide title/s and author/s.	-

Reviewer (Student number) D10125115 Proposer (Student number) C06473075

How can the definition of the domain/scope of the research being proposed be improved? and why?	Domain clear. Scope unclear at presentation.
What should the proposer include in the literature review? and why?	N/A
Are the gaps found in the literature clear and convincing? How can these be clarified/improved?	Very interesting but not many gaps shown
Is the research problem/question clear? How can it be improved?	Just needs rewording into a question. Otherwise clear
Can you suggest any relevant research paper/s to the proposer? If yes, provide title/s and author/s.	N/A.

Reviewer (Student number) D13128728 Proposer (Student number) C06473075

How can the definition of the domain/scope of the research being proposed be improved? and why?	lot of that more should be in summary section or motivation section
What should the proposer include in the literature review? and why?	
Are the gaps found in the literature clear and convincing? How can these be clarified/improved?	standards in other geographies USA v Europe v Asia ?
Is the research problem/question clear? How can it be improved?	tokenisation. how to quantify if not comparing against other methods?
Can you suggest any relevant research paper/s to the proposer? If yes, provide title/s and author/s.	different global standards.

Reviewer (Student number) D08117802 Proposer (Student number) C06473075

How can the definition of the domain/scope of the research being proposed be improved? and why?	Domain very clear Scope well defined
What should the proposer include in the literature review? and why?	Interesting topic, very well explained
Are the gaps found in the literature clear and convincing? How can these be clarified/improved?	I'm convinced but I have no background or real knowledge of fraud and fraud detection
Is the research problem/question clear? How can it be improved?	Clear statement
Can you suggest any relevant research paper/s to the proposer? If yes, provide title/s and author/s.	N/A

D14124209

Reviewer (Student number) ~~D13123574~~ Proposer (Student number) C06473075

How can the definition of the domain/scope of the research being proposed be improved? and why?	N/A
What should the proposer include in the literature review? and why?	N/A
Are the gaps found in the literature clear and convincing? How can these be clarified/improved?	Yes. N/A
Is the research problem/question clear? How can it be improved?	Yes. But it seems a big/wide topic for a MSc thesis. Maybe it would be ok to restrict the scope to events or experiment
Can you suggest any relevant research paper/s to the proposer? If yes, provide title/s and author/s.	N/A

Reviewer (Student number) D13122545 Proposer (Student number) C06473075

How can the definition of the domain/scope of the research being proposed be improved? and why?	SEPERATE METHODS FROM TYPES AS BUNCHED AT MOMENT
What should the proposer include in the literature review? and why?	WHAT IS PCI DSS, EMV, CNP, AND TOKENISATION? EXPLAIN THESE
Are the gaps found in the literature clear and convincing? How can these be clarified/improved?	GOOD RANGE OF GAPS TO WORK WITH
Is the research problem/question clear? How can it be improved?	CLARIFY "EFFECTIVENESS OF TOKENISATION"
Can you suggest any relevant research paper/s to the proposer? If yes, provide title/s and author/s.	N/A

Reviewer (Student number) C0357827 Proposer (Student number) C06473075

How can the definition of the domain/scope of the research being proposed be improved? and why?	Well explained
What should the proposer include in the literature review? and why?	√ Great
Are the gaps found in the literature clear and convincing? How can these be clarified/improved?	Good interesting gaps
Is the research problem/question clear? How can it be improved?	0/5 for no ? ☺ I didn't understand the statement
Can you suggest any relevant research paper/s to the proposer? If yes, provide title/s and author/s.	No, sorry

Reviewer (Student number) D13128708 Proposer (Student number) C06473075

How can the definition of the domain/scope of the research being proposed be improved? and why?	Clear.
What should the proposer include in the literature review? and why?	More papers, and trends in Findings between papers, would be more compelling than Findings from single papers.
Are the gaps found in the literature clear and convincing? How can these be clarified/improved?	Seemed clear.
Is the research problem/question clear? How can it be improved?	Not a clear question.
Can you suggest any relevant research paper/s to the proposer? If yes, provide title/s and author/s.	Don't know any.

Reviewer (Student number) D10121831 Proposer (Student number) C06473075

How can the definition of the domain/scope of the research being proposed be improved? and why?	Good Clear Problem definition Scope and domain.
What should the proposer include in the literature review? and why?	Good literature Review.
Are the gaps found in the literature clear and convincing? How can these be clarified/improved?	-Difficult subject matter if data is directly hard to get.
Is the research problem/question clear? How can it be improved?	Research Problem needs to be scoped in. Clarify effectiveness
Can you suggest any relevant research paper/s to the proposer? If yes, provide title/s and author/s.	N/A

Reviewer (Student number) D13122773 Proposer (Student number) C06473075

How can the definition of the domain/scope of the research being proposed be improved? and why?	Very good.
What should the proposer include in the literature review? and why?	Include the ~~definition~~ criteria of "effectiveness" in q.
Are the gaps found in the literature clear and convincing? How can these be clarified/improved?	
Is the research problem/question clear? How can it be improved?	very clear, but focus as question.
Can you suggest any relevant research paper/s to the proposer? If yes, provide title/s and author/s.	

Reviewer (Student number) D14124953 Proposer (Student number) C06473075

How can the definition of the domain/scope of the research being proposed be improved? and why?	Domain & Scope well defined
What should the proposer include in the literature review? and why?	Maybe include some search terms
Are the gaps found in the literature clear and convincing? How can these be clarified/improved?	Literature gaps seem a bit sparse
Is the research problem/question clear? How can it be improved?	needs a question mark what is effectiveness? what is tokenisation?
Can you suggest any relevant research paper/s to the proposer? If yes, provide title/s and author/s.	

Reviewer (Student number) d15123771 Proposer (Student number) C06473075

How can the definition of the domain/scope of the research being proposed be improved? and why?	Well structured.
What should the proposer include in the literature review? and why?	/
Are the gaps found in the literature clear and convincing? How can these be clarified/improved?	/
Is the research problem/question clear? How can it be improved?	What is effectiveness?
Can you suggest any relevant research paper/s to the proposer? If yes, provide title/s and author/s.	No.

Reviewer (Student number) C05416591 Proposer (Student number) C06473075

How can the definition of the domain/scope of the research being proposed be improved? and why?	Maybe narrow this down a bit as you move forward.
What should the proposer include in the literature review? and why?	You spoke alot in relation to the USA, is your research specific to the USA?
Are the gaps found in the literature clear and convincing? How can these be clarified/improved?	Yes, clear gaps
Is the research problem/question clear? How can it be improved?	Very clear and great idea
Can you suggest any relevant research paper/s to the proposer? If yes, provide title/s and author/s.	No

Reviewer (Student number) D08121521O Proposer (Student number) C06473075

How can the definition of the domain/scope of the research being proposed be improved? and why?	Clear.
What should the proposer include in the literature review? and why?	Interesting.
Are the gaps found in the literature clear and convincing? How can these be clarified/improved?	—
Is the research problem/question clear? How can it be improved?	What do you mean about tokenisation?
Can you suggest any relevant research paper/s to the proposer? If yes, provide title/s and author/s.	—

C06473075

Reviewer (Student number) C03780210 Proposer (Student number) [redacted]

How can the definition of the domain/scope of the research being proposed be improved? and why?	where will you be getting the data?
What should the proposer include in the literature review? and why?	Tokenisation. not sure what this means
Are the gaps found in the literature clear and convincing? How can these be clarified/improved?	Gaps are clear and relevant
Is the research problem/question clear? How can it be improved?	Could possibly be a little more precise but seems very interesting
Can you suggest any relevant research paper/s to the proposer? If yes, provide title/s and author/s.	None, sorry.

("To what extent can tokenisation mitigate "card not present" fraud in a global EMV environment?" in summary

Reviewer (Student number) C99391341 Proposer (Student number) C06473075

How can the definition of the domain/scope of the research being proposed be improved? and why?	Domain - "evolving domain of Card fraud" Very clear & includes a lot.
What should the proposer include in the literature review? and why?	Seems comprehensive
Are the gaps found in the literature clear and convincing? How can these be clarified/improved?	How will you get data for this - will banks etc. release raw data or is it kept private as ~~p~~ personal data?
Is the research problem/question clear? How can it be improved?	How to define "effectiveness"? Not a question!
Can you suggest any relevant research paper/s to the proposer? If yes, provide title/s and author/s.	No, sorry!

Reviewer (Student number) D13122574 Proposer (Student number) C06473075

How can the definition of the domain/scope of the research being proposed be improved? and why?	thats clear
What should the proposer include in the literature review? and why?	don't know
Are the gaps found in the literature clear and convincing? How can these be clarified/improved?	possibly define tokenisation ~~as~~ for those not familiar.
Is the research problem/question clear? How can it be improved?	define effectiveness?
Can you suggest any relevant research paper/s to the proposer? If yes, provide title/s and author/s.	N/A

C06473075	Which method, or combination of methods, would minimise the number of false positives found during fraud detection of card not present fraud?	Two methods of fraud detection can be combined to effectively identify fraudulent transactions to an industry acceptable standard while minimising the number of legitimate transactions that are often incorrectly identified as fraudulent.
	Question is clear but context is vague - data from what location, what period, ~~date~~ what type of transaction etc.	Hypothesis: Combining [Method A] and [B] decreases the false positive rate when detecting fraud
C99391341	What are the factors that influence intake of applications to the Property Registration Authority, and can these features be used to build models to predict future intake?	That the future intake of applications in the Property Registration Authority can be accurately predicted using both historical intake data and external data relating to the property market and state of the economy.
	Clear question.	Hypothesis: Historical intake data, and current property and economic data can predict future intake of applications... The future volume? Per annum, per quarter, per month? Is this limited to Dublin, Non-Dublin, All-Island?

C06473075	Which method, or combination of methods, would minimise the number of false positives found during fraud detection of card not present fraud?	Two methods of fraud detection can be combined to effectively identify fraudulent transactions to an industry acceptable standard while minimising the number of legitimate transactions that are often incorrectly identified as fraudulent.
	OK	EXAMPLES ARE?
C99391341	What are the factors that influence intake of applications to the Property Registration Authority, and can these features be used to build models to predict future intake?	That the future intake of applications in the Property Registration Authority can be accurately predicted using both historical intake data and external data relating to the property market and state of the economy.
	OK	TO WHAT % AGE DESIRED?

C06473075	Which method, or combination of methods, would minimise the number of false positives found during fraud detection of card not present fraud?	Two methods of fraud detection can be combined to effectively identify fraudulent transactions to an industry acceptable standard, while minimising the number of legitimate transactions that are often incorrectly identified as fraudulent.
	Maybe the wording could be changed slightly to make more clear not clear → clear	very clear ~~SCORES~~ good
C99391341	What are the factors that influence intake of applications to the Property Registration Authority, and can these features be used to build models to predict future intake?	That the future intake of applications in the Property Registration Authority can be accurately predicted using both historical intake data and external data relating to the property market and state of the economy.
	Very clear	very clear → do you have to say how you intend to prove eg 'using Hacking leaking techniques'?

C06473075	Which method, or combination of methods, would minimise the number of false positives found during fraud detection of card not present fraud?	Multiple Two methods of fraud detection can be combined to effectively identify fraudulent transactions to an industry acceptable standard while minimising the number of legitimate transactions that are often incorrectly identified as fraudulent. lowering
	Use ensemble methods to minimise false positives returned during fraud detection of CNP fraud	Ensemble methods ~~per can~~ lower the number of false positives of the overall model
C99391341	What are the factors that influence intake of applications to the Property Registration Authority, and can these features be used to build models to predict future intake?	if you say only future then you cannot test your hypothesis define accurately That the ~~future~~ intake of applications in the Property Registration Authority can be accurately predicted using both historical intake data and external data relating to the property market and state of the economy. [or leave it out]
	Perhaps rank these factors [include list in the question]	

C06473075	Which method, or combination of methods, would minimise the number of false positives found during fraud detection of card not present fraud?	Two methods of fraud detection can be combined to effectively identify fraudulent transactions to an industry acceptable standard while minimising the number of legitimate transactions that are often incorrectly identified as fraudulent.
		clear
C99391341	What are the factors that influence intake of applications to the Property Registration Authority, and can these features be used to build models to predict future intake?	That the future intake of applications in the Property Registration Authority can be accurately predicted using both historical intake data and external data relating to the property market and state of the economy.
	clear	clear

C06473075	Which method, or combination of methods, would minimise the number of false positives found during fraud detection of card not present fraud?	Two methods of fraud detection can be combined to effectively identify fraudulent transactions to an industry acceptable standard while minimising the number of legitimate transactions that are often incorrectly identified as fraudulent.
	good Which method(s) would minimise the number of false +ves found during fraud detection of card not present found ?	Name the methods x + y can be combined to effectively identify fraud and transactions ~~too~~ better than using x + y independently.
C99391341	What are the factors that influence intake of applications to the Property Registration Authority, and can these features be used to build models to predict future intake?	That the future intake of applications in the Property Registration Authority can be accurately predicted using both historical intake data and external data relating to the property market and state of the economy.
	good	

C06473075	Which method, or combination of methods, would minimise the number of false positives found during fraud detection of card not present fraud?	Two methods of fraud detection can be combined to effectively identify fraudulent transactions to an industry acceptable standard while minimising the number of legitimate transactions that are often incorrectly identified as fraudulent.
		do you have access to actual current data, and will you be able to test your methods on the same data set.
C99391341	What are the factors that influence intake of applications to the Property Registration Authority, and can these features be used to build models to predict future intake?	That the future intake of applications in the Property Registration Authority can be accurately predicted using both historical intake data and external data relating to the property market and state of the economy.
	Should you say more about the kind of model that you want to build?	

C06473075	Which method, or combination of methods, would minimise the number of false positives found during fraud detection of card not present fraud?	Two methods of fraud detection can be combined to effectively identify fraudulent transactions to an industry acceptable standard while minimising the number of legitimate transactions that are often incorrectly identified as fraudulent.
	MAYBE NEED TO SAY IF THESE ARE MACHINE LEARNING METHODS	& SAY HOW CORRECT/INCORRECT WILL BE QUANTIFIED
C99391341	What are the factors that influence intake of applications to the Property Registration Authority, and can these features be used to build models to predict future intake?	That the future intake of applications in the Property Registration Authority can be accurately predicted using both historical intake data and external data relating to the property market and state of the economy.
	SEEMS BROAD QUESTION; ECONOMIC FACTORS? TECHNICAL ISSUES? OK! IT'S PRETTY CLEAR HERE	

Student number	[illegible] C06473075

	Strongly disagree	disagree	undecided	agree	Strongly disagree
Research question identified is too broad or vague		✓			
Hypotheses are appropriate for tackling the research question				✓	
Research objectives have been identified for each hypothesis				✓	
Research methods have been identified for each hypothesis				✓	
Research plan provided will not answer the question or uses other inappropriate research methodologies		✓			

Extra feedback:	

Student number	C06473075

	Strongly disagree	disagree	undecided	agree	Strongly disagree
Research question identified is too broad or vague		-			
Hypotheses are appropriate for tackling the research question					-
Research objectives have been identified for each hypothesis					-
Research methods have been identified for each hypothesis					-
Research plan provided will not answer the question or uses other inappropriate research methodologies	-				

Extra feedback:	

Student number	[illegible]

	Strongly disagree	disagree	undecided	agree	Strongly disagree
Research question identified is too broad or vague		X			
Hypotheses are appropriate for tackling the research question				X	
Research objectives have been identified for each hypothesis				X	
Research methods have been identified for each hypothesis				X	
Research plan provided will not answer the question or uses other inappropriate research methodologies		X			

Extra feedback:	Looks good

Student number	C06

	Strongly disagree	disagree	undecided	agree	Strongly disagree
Research question identified is too broad or vague		X			
Hypotheses are appropriate for tackling the research question				X	
Research objectives have been identified for each hypothesis				X	
Research methods have been identified for each hypothesis				X	
Research plan provided will not answer the question or uses other inappropriate research methodologies		X			

Extra feedback:	No suggestions ☺

Student number	C06473075

	Strongly disagree	disagree	undecided	agree	Strongly disagree
Research question identified is too broad or vague		✓			
Hypotheses are appropriate for tackling the research question				✓	
Research objectives have been identified for each hypothesis				✓	
Research methods have been identified for each hypothesis				✓	
Research plan provided will not answer the question or uses other inappropriate research methodologies		✓			

Extra feedback:	

Student number	C06473075

	Strongly disagree	disagree	undecided	agree	Strongly disagree
Research question identified is too broad or vague		✓			
Hypotheses are appropriate for tackling the research question				✓	
Research objectives have been identified for each hypothesis				✓	
Research methods have been identified for each hypothesis				✓	
Research plan provided will not answer the question or uses other inappropriate research methodologies		✓			

Extra feedback:	

Student number

C06473075	Strongly disagree	disagree	undecided	agree	Strongly disagree
Research question identified is too broad or vague			X		
Hypotheses are appropriate for tackling the research question			X		
Research objectives have been identified for each hypothesis				X	
Research methods have been identified for each hypothesis				X	
Research plan provided will not answer the question or uses other inappropriate research methodologies			X		

Extra feedback:	What data will be used?

Student number C0647 3075

	Strongly disagree	disagree	undecided	agree	Strongly disagree
Research question identified is too broad or vague		✓			
Hypotheses are appropriate for tackling the research question				✓	
Research objectives have been identified for each hypothesis				✓	
Research methods have been identified for each hypothesis				✓	
Research plan provided will not answer the question or uses other inappropriate research methodologies		✓			

Extra feedback:	

Student number C0647 3075

	Strongly disagree	disagree	undecided	agree	Strongly disagree
Research question identified is too broad or vague			✓		
Hypotheses are appropriate for tackling the research question				✓	
Research objectives have been identified for each hypothesis				✓	
Research methods have been identified for each hypothesis				✓	
Research plan provided will not answer the question or uses other inappropriate research methodologies		✓			

Extra feedback:	No suggestions

Student number	C06473075

	Strongly disagree	disagree	undecided	agree	Strongly disagree
Research question identified is too broad or vague			X		
Hypotheses are appropriate for tackling the research question				X	
Research objectives have been identified for each hypothesis				X	
Research methods have been identified for each hypothesis			X		
Research plan provided will not answer the question or uses other inappropriate research methodologies				X	

Extra feedback:	big Data platform is a vague term in the research question provide more details about Accuracy

Student number	

	Strongly disagree	disagree	undecided	agree	Strongly disagree
Research question identified is too broad or vague	✓				
Hypotheses are appropriate for tackling the research question				✓	
Research objectives have been identified for each hypothesis				✓	
Research methods have been identified for each hypothesis				✓	
Research plan provided will not answer the question or uses other inappropriate research methodologies		✓			

Extra feedback:	

Student number	CO

	Strongly disagree	disagree	undecided	agree	Strongly disagree
Research question identified is too broad or vague				✓	
Hypotheses are appropriate for tackling the research question				✓	
Research objectives have been identified for each hypothesis				✓	
Research methods have been identified for each hypothesis				✓	
Research plan provided will not answer the question or uses other inappropriate research methodologies		✓			

Extra feedback:	* RQ - "Alleviate ..." → to what degree? * Just to avoid bias - you shouldn't be hopeful that it will be better; you have to be objective!

Student number	

	Strongly disagree	disagree	undecided	agree	Strongly disagree
Research question identified is too broad or vague		✓			
Hypotheses are appropriate for tackling the research question				✓	
Research objectives have been identified for each hypothesis				✓	
Research methods have been identified for each hypothesis				✓	
Research plan provided will not answer the question or uses other inappropriate research methodologies		✓			

Extra feedback:	

Student number	C0547 3075

	Strongly disagree	disagree	undecided	agree	Strongly disagree
Research question identified is too broad or vague	✓				
Hypotheses are appropriate for tackling the research question				✓	
Research objectives have been identified for each hypothesis				✓	
Research methods have been identified for each hypothesis				✓	
Research plan provided will not answer the question or uses other inappropriate research methodologies	✓				

Extra feedback:	Sauds good ☺

Student number	C064 7307 5

	Strongly disagree	disagree	undecided	agree	Strongly disagree
Research question identified is too broad or vague		✓			
Hypotheses are appropriate for tackling the research question				✓	
Research objectives have been identified for each hypothesis				✓	
Research methods have been identified for each hypothesis				✓	
Research plan provided will not answer the question or uses other inappropriate research methodologies		✓			

Extra feedback:	

Student number	

	Strongly disagree	disagree	undecided	agree	Strongly disagree
Research question identified is too broad or vague			✓		
Hypotheses are appropriate for tackling the research question				✓	
Research objectives have been identified for each hypothesis				✓	
Research methods have been identified for each hypothesis				✓	
Research plan provided will not answer the question or uses other inappropriate research methodologies		✓			

Extra feedback:	Not sure about the Research Question being too vague to a non-expert. Everything else seems great.

Student number	C0547307S

	Strongly disagree	disagree	undecided	agree	Strongly disagree
Research question identified is too broad or vague		X			
Hypotheses are appropriate for tackling the research question				✓	
Research objectives have been identified for each hypothesis				✓	
Research methods have been identified for each hypothesis				✓	
Research plan provided will not answer the question or uses other inappropriate research methodologies		✓			

Extra feedback:	

Student number	C06473075

	Strongly disagree	disagree	undecided	agree	Strongly disagree
Research question identified is too broad or vague				✓	
Hypotheses are appropriate for tackling the research question			✓		
Research objectives have been identified for each hypothesis		✓			
Research methods have been identified for each hypothesis				✓	
Research plan provided will not answer the question or uses other inappropriate research methodologies			✓		

Extra feedback:	ADD OBJECTIVES

Student number	C06473075 NAOMI

	Strongly disagree	disagree	undecided	agree	Strongly disagree
Research question identified is too broad or vague		✓			
Hypotheses are appropriate for tackling the research question				✓	
Research objectives have been identified for each hypothesis				✓	
Research methods have been identified for each hypothesis				✓	
Research plan provided will not answer the question or uses other inappropriate research methodologies		✓			

Extra feedback:	How specifically are you going to compare performance, especially as are you going to apply a statistical test.

Student number	C06473075

	Strongly disagree	disagree	undecided	agree	Strongly disagree
Research question identified is too broad or vague					
Hypotheses are appropriate for tackling the research question					
Research objectives have been identified for each hypothesis					
Research methods have been identified for each hypothesis					
Research plan provided will not answer the question or uses other inappropriate research methodologies					

Extra feedback:	Assume any differences in average performance will be not statistically relevant

Student number	C06473075

	Strongly disagree	disagree	undecided	agree	Strongly disagree
Research question identified is too broad or vague		X			
Hypotheses are appropriate for tackling the research question				X	
Research objectives have been identified for each hypothesis				X	
Research methods have been identified for each hypothesis				X	
Research plan provided will not answer the question or uses other inappropriate research methodologies		X			

Extra feedback:	