# A picture containing clock Description automatically generatedSchool of Information Technology Department of Computer Science (MIT 807)

**Project Proposal: R. Domingues st.no. 93398761**

Mini-Dissertation Title: Towards Optimising Topic Classification of Documents for Forensic Investigators

# Introduction

Computer forensic analysts

The introduction provides the reader with sufficient information on the field of study to allow for an appreciation of the proposed research i.e. it places the project in context with what is known. This section can also form the basis for the introduction of the research proposal/dissertation.

# Aim

A clear statement of the general aims of the project plus a set of objectives that are potentially achievable.

# Hypotheses and Questions

The hypothesis is a tentative theory about the natural world that can be tested by further investigation, whilst the questions allow for more focussed or specific attention to particular aspects of the project. It is essential that the hypotheses and/or questions are clearly and unambiguously stated. Since it is these statements that are going to guide the research through the practical portion of the project and will also assume central roles in the written mini-dissertation it is important to consider their wording and composition very carefully. Many students find this a difficult task.

# Methodology

This section varies with each discipline. In essence, the section provides a detailed description of what material is to be used, what experimentation is to be conducted, what data are to be collected and how the data are to be analysed. It is worth spending time on this section as it will give you a clear indication of what work you will be conducting and assists you in planning

the project. Consideration must also be given to the following: whether the material is readily available; whether permission is required to gain access to the material; whether ethics clearance is required. Solutions to potential problems must also be considered.

# Goals and processes

This section outline details of what work needs to be done and when. It is essential that when this plan is prepared, you are realistic about what can be achieved at any given time. You will find that you often underestimate how long it will take to achieve a particular task, check your timings with your supervisor or other senior postgraduates in your field.

As is common in the industry and the broader workplace, many supervisors also ask for a budget to be prepared for the project (If funding is required for the project).

A proposal document is considered by the Department of Computer postgraduate committee or supervisor. Adjustments and corrections to the document may be recommended by the committee or by the supervisor. Once the project proposal has been accepted by the committee or supervisor, the document is signed by the candidate, supervisor(s), coordinator and Head of the Department. The signed document is submitted to the Department Secretary and a formal acknowledgement by the candidate to undertake the research under the guidance of the supervisor(s).

The proposal document must be submitted to the MIT coordinator within a specified period of the original registration. The students must finalise their research topic within the first year of their registration to avoid a further deal in finishing the program within the stipulated period.

# Evaluation of Research Questions

You should have a plan for testing your system when it is complete. Work this out now; everything will be wasted if you finish your implementation but cannot evaluate your “advance” convincingly. Indicate the interpretation and conclusions that you will place upon the results. What difference will they make? Indicate the implications of your research for current theory and practice.

# Work Detail

Decide on the stages of the project and the dependencies between them. Compile a project plan.

* Timeline, including Gantt chart. Use specific dates so that you finish on time.
* Resources required (equipment, people, special software etc)
* Deliverables
* Milestones (which should refer to the Timeline)

# Bibliography

List the main sources on which your research will be based. In the proposal, we want a preliminary outline of the key works. All studies must be properly cited. As you progress in writing your research proposal have to show that you have read the relevant papers and books and understand the field. You should show that you know which important contributions are and how they are related and may be grouped. You should know where the concepts you use were first described.

In digital forensic investigations [1], data is seized from computers, cellphones, cloud services, email accounts etc. The data is then copied and analyzed by dedicated investigators. The investigators need to deal with large volumes of data which includes pictures, video, voice recordings, emails, documents etc. Various commercial forensic tools exist that the investigators use to search the digital data streams to extract and catalog the information into these categories [1]. With a background into what is being investigated, the forensic analysts need to work through the categorized data to find evidence. A big part of the evidence is in the form of text (e.g. emails, documents), and a lot of time is initially wasted in sifting through the text documents to determine which documents to focus on [1]. For example, investigators can initially ignore all emails that come from newsletters and documents of a general nature (e.g. product information). Investigators typically want to initially focus on interpreting financial related documents, emails concerning specific topics relevant to the investigation etc.

A dedicated investigator may be from a social sciences background and has to analyze seized communications between suspects. Communications phrases and terms may occur which the investigator, as a non-subject matter expert, may not be familiar with (e.g. components of a technical system). The investigator can better interpret the information if the investigator knows the sentence is related to a topic (e.g. the designs of the Boeing 777) [2]. In the described scenario, all related documents are available to the investigator, but the volumes are large. The investigator can in theory conduct a keyword search, but proper keyword selection is an art better practiced by a subject matter expert.

These two described scenarios can be aided by performing topic analysis and classifying the source documents in terms of topics [1] [2]. In the first example of finding the documents to focus on, topic analysis needs to produce good summaries of the topics covered in the various documents. A skilled investigator will use the topics and match them with the evidence they suspect to find, and then first focus on the related documents. We will refer to this as application 1 in this proposal.

In the second example presented here, the identified phrases will indicate the topics related to the phrase topic and map the topic back to a general topic (e.g. Turbine blades maps back to airplane engines, and Rolls Royce model XT220 maps back to the Boeing 777 engine). We will refer to this example as application 2 in this proposal.

The rest of this proposal is structured as follows. Section II introduces the problem with current approaches and formulates a problem statement. The problem statement is then used to formulate a research question for the intended research. Research sub-questions are formulated which, when addressed, will address the research question.

Section III analyzes related research, points out the shortcomings, as well as useful techniques that can be applied towards answering the research question. Section IV discuss the proposed methodology that will be used to address the research questions. Section V provides a high-level planning to establish timelines and goals in order to complete the research in a realistic timeframe.

# Problem Statement

In the literature we surveyed, most methods focus on clustering as a means of extracting topics from documents. Very little of the literature addresses the question of whether the topic analysis produces relevant results that would be useful to the forensic analysists. Most literature only demonstrate that the tools in text mining may be useful in application 1 described in this proposal, while application 2 described in this proposal is not addressed.

Some literature consider optimization in terms of the technical tools used during the clustering process (e.g. clustering method). The optimization techniques are often of a theoretical nature, e.g. showing that k-means clustering performs better than other clustering techniques under certain conditions. There are many parameters that can be optimized in topic extraction on a corpus of documents, some of which is not related to the clustering algorithm. Deciding on which parameters to optimize is dependent on what output will be useful to the forensic analyst. This leads us to the following problem statement:

* *Choosing the optimal parameters for topic analysis is not only dependent on the data, but also on the desired outcome for the application that the forensic analyst desires.*

From the above problem statement, we can formulate the following research question that we want to address in the proposed research:

* *Which parameters in an application of clustering to extract topics from documents can be adjusted to optimize the resultant topic outputs for application 1 and 2 described in this proposal?*

In order to address the main research question, we formulate the following sub-questions that needs to be addressed.

* *What is the desired outcome from topic analysis that will assist the investigators from their perspective to make the topic analysis tools useful in a real investigation?*
* *What are the parameters of the various categories of data they deal with (e.g. document size, sentence lengths, document structures)?*
* *What are the parameters in a topic analysis using clustering that can be optimized?*
* *Which of the parameters are most important to produce the best results in terms of the desired outcome for the investigators?*

# Literature Survey

In this section we only address the most important literature that we surveyed. We will cover what was previously attempted, what optimizations were tested, and what other possible techniques exist which was not attempted in the context of forensic investigations. We will end off with the gaps we believe exists in the literature that needs to be addressed.

## Previous Work

In [3] the challenges of analyzing big volumes of data in forensic analysis is discussed. A two-step process is then proposed for topic extraction from documents. The first step entails the text extraction, and the second step utilize clustering to group topics. Background is also provided on the actual extraction process from a technical point of view. For example, extracting text data from PDF documents vs Word documents is a challenge. In [3] reference is also made to partial documents through recovery of deleted files. For our study, we will only consider documents that are fully intact. The text extraction is vectorized per document in [3] and K-means clustering is applied to the vectorized documents to determine distances between the documents. Clusters on the distances are used to represent topics. In the examples in [3] the Enron dataset is analyzed for three email addresses. The output consists of words identified as topic related. From our perspective, this is keyword extraction, and not necessarily useful for summarizing document topics. No mention is made of benchmarking the algorithm in practice.

Clustering and sentence extraction is used in [4] to generate summaries of documents. A measure of sentence similarity is used to define a distance function between sentences in a document. Optimizing the clustering approach, a particle swarm optimization algorithm is used to define clusters. Ranking of sentences is then used to summarize the documents. The experiments are performed against datasets which comes from different categories. A strong point of the approach followed in [4] is, although more text is used, sentences make more sense to humans than words. The idea of ranking of sentences is important from our perspective.

A recent literature survey [5] summarizes techniques employed in text clustering for topic extraction, including optimizations that were attempted. It is a nice overview of attempts and the results of the attempts. Proper motivation is provided in [5] that current forensic tools rely on the investigator to know which search terms to use to identify the documents of interest. A good overview and comparison of all the techniques are provided.

## Optimizations and Measurements

In [3] no optimizations were attempted. The measurements were subjective based on knowledge of the emails and some background information available. In [4] the technique of applying particle swarm optimization is described as an optimization of finding parameters. This type of optimization has little to do with improving accuracy and not relevant for our purposes.

The work summarized in [5] contains many optimization techniques. One that is noteworthy is the clustering algorithms employed. K-means clustering is commonly used, and even clustering ensembles were investigated in the literature. We believe as a technical optimization research are, there are newer techniques that can be compared, specifically KNN Clustering. In [6] KNN Clustering and KNN Clustering ensembles are investigated in the general setting. This is a new clustering technique and is not covered in [5]. Also, a careful analysis of the clustering ensemble proposal in [6] reveals that there are many ways to construct clustering ensembles.

The first text we could find that attempts to match the topic extraction to forensic analysis is [7]. In Section II of [7] a requirements analysis is performed and in Section III an ontology for investigation is developed. Section IV then discusses text analysis in terms of this ontology. It is clear that the authors of [7] improved their work by shifting focus. The unpublished paper [8] took a different approach which was closer to the approaches discussed up to now. In our opinion, the approach followed in [7] to develop a better understanding of the application environments will lead to refined text analysis models. These refined models are applicable to specific domains, and allow for better measurements of success.

## Other Techniques

The PhD thesis [9] investigates text analysis in security classification (e.g. Restricted, Confidential, Secret) problems. In the documents that were experimented on, paragraphs are individually classified. The thesis provides a good overview of automated text classification, including clustering techniques applied in text classification. The algorithm introduced operates on paragraph level instead of sentence level or document level. Optimization through topic model pruning is also explored. Though [9] is not literature on forensic analysis, the concepts of paragraph level classification as well as optimization through topic pruning can be applied in the research we are interested in. To our knowledge there are not papers addressing this in the context of forensic investigations.

## Summary of Identified Gaps

We summarize the gaps as follow. The literature focus on what is believed to be useful in terms of topic analysis and classification for forensic investigators. There is no mapping to what a forensic investigator requires as the output. The specific shortcoming is a description of the requirements. Some of the work performed in [7] may address application 1 described in this proposal, but not necessarily application 2. The usefulness of term-based summaries vs sentence-based summaries is therefore unknown.

There is no comparison between term extraction, sentence level extraction and paragraph level extraction of topics to form summaries. In fact, paragraph level topic analysis models are not even considered in texts dealing with topic analysis in forensic investigations.

Finally, a new clustering technique namely KNN Clustering and various options for ensemble clustering models was not fully analyzed in the literature.

# Methodology

Our approach to this research will be practical and will require a few iterations to refine. The methodology proposed here is designed to address the sub-questions as set out in Section II.

Step I. A detailed analysis of the requirements for application 1 and application 2 will be documented. The requirements will be discussed with investigators working in the two environments [1] [2]. Further refinement of the requirements will take place through iterative demonstrations of the output generated by implemented models (discussed next).

Step II. Suitable models will be selected to perform the topic analysis on various test datasets. Basic models for term-based, sentence-based and paragraph-based approaches to topic analysis will be constructed. The output of the topic analysis or summaries will be as closely matched to what was documented in Step I above. No specific optimizations will be attempted. This will form the baseline of later optimization. It will also serve as a subjective measurement of the requirements. Demonstrations in the two environments [1] [2] with discussions will aid refinement of an understanding of the requirements. The Enron email dataset used by [3] will be used as data, as well as the leaked cables used by [9].

Step III. With a better understanding of the requirements, and a baseline established in Step II, the various parameters will be optimized and the effects on the documented requirements will be noted. Specifically, clustering ensembles and KNN Clustering will be evaluated in these models and compared to other clustering algorithm techniques. This step will take the most effort as evaluation will be subjective and the identification of all the parameters will be hard. The outcome of this step must be a well-documented explanation of the effect of the tweaking of parameters on the desired outcomes from Step I.

Step IV. The selected algorithms will be applied on real data in the environments of applications 1 and 2, and parameters will be further optimized. Subjective feedback from the investigators will be used for the refinements. All refinements will be properly documented.

Limitations. In this proposal no objective performance evaluations are proposed. This is due to limited access to willing investigators. The purpose of this work is to establish a baseline. Future research can develop more objective measures of performance of the topic analysis when a better understanding of the requirements is well established.

# Planning

We plan to conduct this research over one academic year. The academic year is deemed to start in February and final results must be written up by November. The table below sets the high-level goals and target dates for those goals. In each case the target date is at the end of the month.

|  |  |
| --- | --- |
| **Goal Description** | **Target Completion Date** |
| Prepare documentation and apply for ethical clearance for the final refinement of the algorithms on real data. | February |
| Interviews and finalization of first requirements document (Step I) | February |
| Implementation of concepts models | March |
| Second interview, demonstration of models and refinement of requirements (Step II). | May |
| Identification of parameters for optimization and experimentation (Step III). | July |
| Documentation of parameter optimization. | August |
| Demonstrations to application environments, further refinement on real data. | September |
| Possible development of academic papers to be submitted to conferences. | October |
| Final writeup of results and presentation. | November |

Risks and Mitigation. No explicit time is allocated to documenting results. We believe this can be mitigated by continuous documentation throughout the research project.

The availability of investigators for interviews and participation is difficult to determine in advance. The investigators may also have heavy caseloads which will dictate their availability. Early engagement with investigators (before the start of the academic year) will be required to mitigate this risk.

Ethical clearance might delay the final step of testing on real data. An application can be submitted within this year. If ethical clearance cannot be obtained in time (September), the final steps can be omitted, or the plan will have to be extended into the new year. Working over the holiday season might be required in that case by the researcher, and final submission can take place at the start of the new academic year.

##### References

[1] J. Malan, ‘Personal Communications: Text Mining as an aid in Forensic Analysis.’, Nov-2019.

[2] L. Ellerbeck, ‘Personal Communications: Aids to assist investigators in finding the source documents/topics.’, Nov-2019.

[3] S. Decherchi, S. Tacconi, J. Redi, A. Leoncini, F. Sangiacomo, and R. Zunino, ‘Text Clustering for Digital Forensics Analysis’, in *Computational Intelligence in Security for Information Systems*, vol. 63, Á. Herrero, P. Gastaldo, R. Zunino, and E. Corchado, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 29–36.

[4] R. M. Aliguliyev, ‘Clustering Tehniques and Discrete Particle Swarm Optimization Algorithm for Multi-Document Summarization’, *Comput. Intell.*, vol. 26, no. 4, pp. 420–448, Nov. 2010.

[5] B. Almaslukh, ‘Forensic Analysis using Text Clustering in the Age of Large Volume Data: A Review’, *Int. J. Adv. Comput. Sci. Appl.*, vol. 10, no. 6, 2019.

[6] J. Nordhaug Myhre, K. Øyvind Mikalsen, S. Løkse, and R. Jenssen, ‘Robust clustering using a kNN mode seeking ensemble’, *Pattern Recognit.*, vol. 76, pp. 491–505, Apr. 2018.

[7] M. Spranger and D. Labudde, ‘Towards Establishing an Expert System for Forensic Text Analysis’, *Int. J. Adv. Intell. Syst.*, vol. 7, no. 1 and 2, pp. 247–256, 2014.

[8] M. Spranger and D. Labudde, ‘Semantic Tools for Forensics: Approaches in Forensic Text Analysis’, *Draft Int. J. Adv. Intell. Syst.*, 2013.

[9] K. M. ALZHRANI, ‘Towards Automating Big Texts Security Classifcation’, PhD Thesis, B.S., King Saud University, Computer Science, Saudi Arabia, 2007 M.S., Concordia University, Computer Science, Canada, 2011, University of Colorado Colorado Springs, 2018.