1 Justification, Motivation and benefits

Digital forensic investigation have a big data problem. Without tools that can search the data within a small time frame and provide relevant hits, then forensic investigation cannot examine the evidence in a timely manner. Which in turn can negatively effect the justice system capability of convicting criminals.

The ability for search engines to process large amount of forensic data will be benchmarked in this paper. The benchmark can provide relevant information needed to determine whether or not to invest resources on integrating search engines into the digital forensic investigation process.

2 Planned contributions

The contribution is a framework to evaluate the performance of a selection of search engines and their search functionality in the domain of digital forensics.

3 Related work

4 Keywords

Digital forensics, search engines, benchmarking, open source, recall and precision

5 Problem description

Forensic practitioners in digital forensics have to process a large quantitative of structured and unstructured data. The processing of data have to be reliable, forensically sound and preferably be solved with a low memory and time complexity. Forensic practitioners can use one of many Search Engines (SE) to aid them on this task.

By knowing which SE that is out there, which algorithms they use and their performance, then the forensic practitioners can make a conscious decision on which SE that best aid them on the forensic process.

6 Research question

In table 1, 3 different research question is presented with the variables and groups to be tested.

Table 1: Research questions

Research question 1	
Which index strategy leads to best performance for the search engine	
Variable	index strategies
Group	Selection of SE, SE functionality, forensic data

Research question 2	
How well does search engine functionality perform on forensic data?	
Variable	Time complexity, storage/memory complexity, recall, precision, F measure
Group	Selection of SE, SE functionality, forensic data

Research question 3	
How well does search engines perform on forensic data?	
Variable	Time complexity, storage/memory complexity, recall, precision, F measure
Group	Selection of SE, SE functionality, forensic data

7 Risk analysis

The table 2 is used to reference how severe a risk is with respect to impact and likelihood. The colour red indicates that the risk have to be reduced. With yellow the risk should be reduced. And green is the acceptable level of risk. Below the table I have made a list of the 5 most significant risk elements in my thesis.

Table 2: Risk Table

Impact / Likelihood	**Very Unlikely**	**Remote**	**Seldom**	**Probable**	**Frequent**	**Very Frequent**
Severe	🟥	🟥	🟥	🟥	🟥	🟥
Significant	🟨	🟨	🟥	🟥	🟥	🟥
High	🟩	🟨	🟨	🟥	🟥	🟥
Moderate	🟩	🟩	🟨	🟨	🟥	🟥
Low	🟩	🟩	🟩	🟨	🟨	🟥
Minimal	🟩	🟩	🟩	🟨	🟨	🟨

- Not acquiring the forensic dataset needed for the experiment. Katrin Franke said that the forensic lab could obtain the forensic samples for me. But in case they fail coming though with that in the early stages of my thesis, then I should create a backup dataset.

- 🟥 I would need access to some resources in the forensic lab. To minimize the risk of not getting these resources, I should get a written agreement with key players in the forensic lab and have close communication.
- 🟨 A lot of time might be needed to familiarize myself with the different search engines in order to create my experiment. If this overhead is overwhelming, then it could negatively impact the thesis. I could spend some time in the summer vacation to test these search engines
- 🟨 It takes some time before I get the forensic dataset needed to perform any experiment. A solution to this problem can be to have a small dummy dataset for creating a proof of concept. I would still need the larger dataset, but the dummy dataset would allow me to progress.
- 🟨 Loosing time due to sick days. The best way to avoid that sick days effect the thesis is planning and starting working early.

8 Ethical and legal considerations

There are 3 legal considerations:

- The benchmark experiment can only be performed on search engines with licences that allows benchmarking. This can be managed by only selecting those search engines where benchmarking is allowed.
- The nature of the forensic dataset. The dataset should not contain information that is illegal to store.
- Compliance with written or verbal contracts/agreements with how the forensic lab resources used in the thesis should be handled.

9 Choice of methods

The master thesis will use quantitative methodologies in order to answer the research questions. I would first need to select which search engines to test in my experiment. Then I would need to choose which subset of the search engines filters/search functionality to include in the experiment. Then the selected search engines and search functionality have to be setup/implemented on the test environment. A large forensic dataset have to be acquired, so that the experiment will be run with realistic data types and volume.

To answer the first research question I would need to understand how to implement various indexing strategies. And then test how well these perform in the different search engines. All research questions will require implementation of search engines and search functionality to be benchmarked in the test environment.

The list below is the proposed methodology of how to collect data on recall, precision, F-measurement, time complexity and memory complexity for the experiment. These steps are inspired from the paper [1].

1. A query in one form or another (e.g. filter) will be created using search engine X and search functionality Y, to find some relevant data in the forensic dataset
2. Based on the query and domain knowledge of the dataset, on or more people will decide which documents/data are relevant before the execution of the query statement.
3. Execute the query in the SE (start the search). At this step Memory Complexity (MC) and Time Complexity (TC) should be measured of the algorithms. One possible way to measure this is checking the resource management system on the test environment.
4. Based on the number of actual retrieved documents/data and the number of relevant documents/data we can calculate recall, precision and F-measurement.

collecting these data points should be plausible, as information retrieval systems are often evaluated by the recall and precision metrics. And memory and time complexity of the running process are often tracked by the computer operating system.

10 Feasibility study

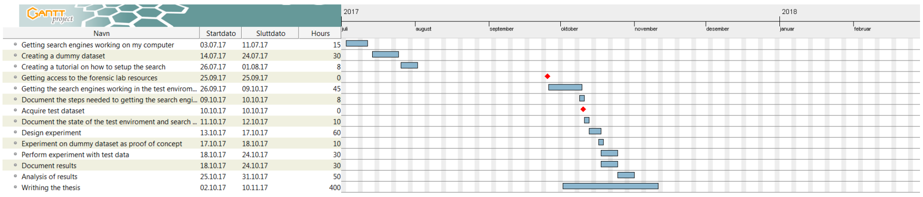
Taking measurements for recall, precision and f-measure for a information retrieval system was done in [1].

Using the documentation and source code for the open source search engine under inspection, it will be easier to understand how to best measure recall, precision, f-measure, time complexity and storage complexity. Data on time and memory complexity can possible also be collected by the resource management system running on the experiment computer environment.

11 Milestones, deliverables and resources

In figure 1 you can see the activities and milestones. The activities can be found under 'navn' and begin date and end date can be found under 'startdato' and 'sluttdato' respectively. The red icon represent a significant event (milestone) in the project. The number of man hours needed to complete an activity can be found under 'hours'. From the chart I can see that the sooner I get access to the forensic resources and test data set, the better. This is because many of the activities depend on them in order progress.

Figure 1: Gant chart (need to zoom in 400%)



Navn	Startdato	Sluttdato	Hours
Getting search engines working on my computer	03.07.17	11.07.17	15
Creating a dummy dataset	14.07.17	24.07.17	30
Creating a tutorial on how to setup the search	26.07.17	01.08.17	8
Getting access to the forensic lab resources	25.09.17	25.09.17	0
Getting the search engines working in the test envirom...	26.09.17	09.10.17	45
Document the steps needed to getting the search engi...	09.10.17	10.10.17	8
Acquire test dataset	10.10.17	10.10.17	0
Document the state of the test enviroment and search ...	11.10.17	12.10.17	10
Design experiment	13.10.17	17.10.17	60
Experiment on dummy dataset as proof of concept	17.10.17	18.10.17	10
Perform experiment with test data	18.10.17	24.10.17	30
Document results	18.10.17	24.10.17	30
Analysis of results	25.10.17	31.10.17	50
Writhing the thesis	02.10.17	10.11.17	400

The deliverables:

- Introduction
- Theory contents
- Description of dataset
- Description on experimental design
- Proof of concept with dummy dataset
- Results section
- Discussion section
- Conclusion
- Abstract

References

[1] Marijan R, Leskovar R. A library's information retrieval system (In)effectiveness: case study. Library Hi Tech. 2015;33(3):369–386.