Search Engines in Digital Forensics

Who? Joachim Hansen

When? March 28, 2017

How to attack the topic I

- My topic is about the performance of search engines on heterogeneous datasets (Structured (data in relational database), Semi-Structured(XML tags) and **Unstructured** (e.g. organization is not decided in advance - book) and the application of search in digital forensics.
  - How to measure the performance of search engines with respect to the needs of forensics practitioners
  - What criteria do I use to decide on which search engines are of interest?
  - Which search algorithms do the search engines use?

How to attack the topic II

- I started to search for scientific articles to get a better understanding of the broader topic
  - ("Abstract":enterprise AND search AND engine), Year: 2014-2017 http://ieeexplore.ieee.org
  - "Enterprise search", in abstract, year: 2014-2017, source type: Scholarly Journals http://search.proquest.com/abicomplete/
  - recordAbstract:(+enterprise +search) , year: 2014-2017 http://dl.acm.org
  - in abstract (Solr OR ElasticSearch) https://arxiv.org
  - in abstract, title and keywords: Enterprise search, year 2014-2017 http://www.sciencedirect.com
  - in abstract:Solr OR ElasticSearch, Scholary journals, year:2014-2017 http://search.proquest.com/
  - in abstract (Solr OR ElasticSearch), year:2014-2017 http://ieeexplore.ieee.org
  - recordAbstract:(Solr Elasticsearch), year:2014-2017 http://dl.acm.org/

How to attack the topic III

- information retrieval unstructured data (general search), year:2014-2017 http://ieeexplore.ieee.org
- "information retrieval" "unstructured data" survey https://scholar.google.no/
- digital forensics information retrieval survey, year:2014-2017, no quoates, no patents https://scholar.google.no/

Performance measures I

- Precision is the fraction of documents that are relevant.
- Recall is the fraction of relevant documents that are retrieved
- F-measure tradeoff between presicon and recall (can add weight to what is the most important).

Performance measures II

- The importance of the measurements precision and recall in Information Retrieval (IR) systems, like Search Engine (SE) depends on the application. In the domain of Digital Forensic Investigation (DFI) precision is more important in the early phases of the forensic investigation, as relevant evidence is vital to guide the process of finding new evidence. At the later stages of the DFI, recall becomes more significant than precision, as Forensic Practitioner (FP) wants all available evidence to build a court case.

Performance measures III

- I was wondering of how you could measure precision and recall of search algorithms (thinking ahead on the master thesis) when these also depend on the quary type, quary quality, how the system interpret the user query, and indexing strategy etc.
- user Query quality depends on their understanding of the quary language and their domain knowledge of the dataset.

Performance measures IV

1. A query in one form or another (e.g. filter) will be created to find some relevant data in the 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.
4. Based on the number of actual retrieved documents/data and the number of relevant documents/data we can calculate recall and precision.
5. Calculate F-measurement which is a measurement for the tradeoff between recall and precision

Search engine criteria I

- What criteria do I use when deciding whether or not a search engine is relevant?
  - Not web search engine - deals mostly with homogeneous data (similar data like HTML) - assumption.
  - Enterprise search engine?
    - The term is loosly used. It can mean search within a enterprise, desktop search, marketing term, search problem - deals with heterogeneous data and advance functionality like security.
  - Open source
  - Currently or recently being updated (in development)
  - deals with unstructured data?
  - No to undocumented search engines?

Search engine criteria II

Name	Algorithms	Dependencies	Last update
Dezi	A_1?,A_2,A_3,S_1 Bookmarked ranking algorithm [**?**],[**?**], [**?**]	D_1,D_2,D_3,D_4 [**?**]	commit 2e82578 28 Nov 2016[**?**]
Apache Solr [**?**]			Solr v 6.4.2: 06 Mar 2017 [**?**]
Sphinx [**?**]			Sphinx 2.3.2-beta 8 Sep 2016 [**?**]
Indri [**?**]			25 Jan 2017 [**?**]
OpenSearchServer [**?**]			commit 5b3e89d 13 Jan 2017 [**?**]
Luwak [**?**]			commit 8336487 6 mar 2017 [**?**]
datafari			commit 72f2bf0 23 mar 2017 [**?**]

Search engine - search algorithms I

- String matching algorithms seems to be omitted from documentation. Could perform a source code inspection? But that could take alot of time.
- Some search engines are poorly documented.
- It might be a naive assumption that the search engines implement all search functionality from the search library API they depend on.
- Look at the application for : Fulltext search, Semantic search, Exploratory search, geospatial search, Fuzzy search, stream search/reverse search, Phonetic search for digital forensics investigation.

Abstract I

Scope:

- An exploration into the application of search in the different phases of Digital Forensics Investigation (DFI).
- Finding open source Enterprise Search Engines (SE) software projects that are not abandoned.
- Considering the major search capabilities of the SE and their strengths and weaknesses with respect to the data type, performance measures like precision and recall and the needs of DFI practitioners.
- Comparison of the string search matching algorithms used by the SE

Abstract II

Forensic practitioners in digital forensics have to process a large quantitative of structured and unstructured data. The processing of data has to be reliable, forensically sound and preferably be solved with a low memory and time complexity. Forensic practitioners (FP) can use a SE to aid them on this task. One encountered limitation of current SE documentation is that string search matching algorithms used by the SE are omitted. Knowledge of the strengths of SE can help FP to decide on which SE to use. The methodology of the paper is to use scientific papers, SE websites (e.g. github), SE software documentation, and SE source code to cover the scope above.