InfoScout

An interactive, entity centric, person search tool

Sean McKeown
School of Computing
Edinburgh Napier University
Edinburgh, Scotland
s.mckeown@napier.ac.uk

Martynas Buivys and Leif Azzopardi
School of Computing Science
University of Glasgow
Glasgow, Scotland
martynas.buivys@gmail.com
Leif.Azzopardi@glasgow.ac.uk

ABSTRACT

Individuals living in highly networked societies publish a large amount of personal, and potentially sensitive, information online. Web investigators can exploit such information for a variety of purposes, such as in background vetting and fraud detection. However, such investigations require a large number of expensive man hours and human effort. This paper describes InfoScout, a search tool which is intended to reduce the time it takes to identify and gather subject centric information on the Web. InfoScout collects relevance feedback information from the investigator in order to rerank search results, allowing the intended information to be discovered more quickly. Users may still direct their search as they see fit, issuing ad-hoc queries and filtering existing results by keywords. Design choices are informed by prior work and industry collaboration.

Keywords

Web Investigation; Open-Source Intelligence; People Searching; Professional Search; Entity Search

1. INTRODUCTION

The World Wide Web is an invaluable resource, providing access to a massive amount of information covering a wide variety of subjects. Ubiquitous access to the Internet has made this information more accessible than ever, while blogging, social media, and communication platforms allow users to be publishers of information, rather than simply consumers. Governments, businesses, and public bodies also contribute to this wealth of information, publishing census data, maps and financial documentation. One way in which these Big Data sources can be exploited is in a variety of Open-Source Intelligence (OSINT)[1] investigations, which make use of publicly available, heterogeneous, information to pursue particular goals. The types of OSINT investigation are varied, ranging from employee vetting [4], competitive intelligence [1], fraud detection [11] and copyright protection [1], to law enforcement [7] and counter-terrorism [3].

Open-Source information, particularly public social media data, is widely adopted for these purposes. Lexis Nexis indicates that 81% of law enforcement professionals in the United States make use of social media platforms as an investigative tool [7], with Facebook, Twitter and Youtube being the most frequented [7]. Similarly, CareerBuilder found that 43% of employers utilise social media to perform background checks on prospective employees, with half of them having rejected candidates as a result of their findings [4].

While such investigations are used in a variety of industries, they are time consuming and fraught with problems. This paper describes an application which is intended to reduce the time it takes for an investigator to find pages which are relevant to a given subject, the target of the investigation, and is based on findings from industry engagement.

2. BACKGROUND

While such types of search are prominent in a variety of industries, the literature in relation to the behaviour of Web investigators is sparse. High level descriptions of processes and techniques can be found in the NATO OSINT Handbook [9], which is set in the context of military like operations. Appel [1] provides a similar overview from the perspective of a professional Web investigator, while providing additional discussion pertaining to industry best practices and ethical considerations.

In order to better understand the requirements of investigators, McKeown et al. [8] conducted a qualitative study which compares and contrasts the behaviour of Web investigators with other search domains. It was found that the investigation is split into two major phases: *(i)* Background Verification and *(ii)* Open-Ended Search. The background verification stage primarily relies on fixed resources, such as resellers of the electoral roll, address information and company directories. The latter focuses more on expanding the knowledge of a subject, largely by way of general Web search engines and social media.

These investigators demonstrated low levels of search expertise, with minimal use of advanced search functionalities. However, a high level of domain expertise was evident. Contextual clues, personal knowledge and experience facilitate the discrimination of search results, as well as the assess-

[1] In this context, the term 'Open-Source' refers to the public availability of information, rather than a software licensing paradigm.

SIGIR '16, July 17 - 21, 2016, Pisa, Italy

ISBN 978-1-4503-4069-4/16/07. . . $15.00
DOI: http://dx.doi.org/10.1145/2911451.2911468

ment of the reliability and quality of pieces of information and sources. Investigators were aware of the notion of a digital footprint, but did not take any technical measures to conceal themselves online. Instead, they employ sensible strategies for avoiding contact with the subject of the investigation, such as carefully considering their interaction with social media profiles.

As a form of domain specific search, such investigations possess properties of exhaustive searches, as in the patent search [6] and E-Discovery domains [2], however, the end goal is not a complete enumeration of all relevant information. High recall is subordinate to high precision, with emphasis on finding the critical pieces of information required to fulfil the purpose of the investigation. The process, then, is more correctly characterised as a form of exploratory search.

However, such investigations are not without difficulty, with issues being highlighted in:

i Person disambiguation and name variants

ii Underutilization of technological solutions for case management and searching

iii Corroboration of findings and the reliability of evidence

iv Time constraints associated with the large volumes of data available on the Web.

2.1 Existing Tools

While investigators in McKeown et al. [8] typically only made use of a popular Web browser, general Web search engines and word processors to carry out the investigation, there are products available for such purposes.

Copernic Agent[2] is meta-search engine which allows search queries to be disseminated to a wide variety of sources for OSINT information gathering. External search engines and resources are categorised thematically, allowing the user to quickly execute searches for a given type of information. However, as of January 2014, this product is no longer supported, with many of the underlying calls to search engines being non-functional at the time of writing.

Maltego[3] is a proprietary OSINT and forensics tool which focuses on allowing the user to determine the links between entities. Maltego has a broad scope and can be used for a variety of OSINT purposes, providing the ability to visualise and discover real world links between entities in social or physical networks. This is achieved by combining meta-search functionality with an underlying graph, allowing the user to explore common nodes or expand the graph to new entities. The search interface differs greatly from that of standard Web search engines, with most actions populating, or operating, on the graph directly. As a result, non-technical users may find this tool difficult to use.

Oryon C Portable[4] is a portable Web browser based on Chromium which is pre-installed with a variety of add-ons and bookmarks for OSINT investigations. The default homepage provides tabs to select common search types, such as Web, Images, Blogs and Maps, while the bookmarks and other features are designed to give an investigator fast access to common resources. Specific resources are searched individually, and the browser itself is geared towards protecting the privacy of the user.

Person searching makes up a substantial portion of all Web searches [10, 5], and, as a result, several people search engines exist, such as *Pipl*[5] and *Spokeo*[6]. Such search engines typically behave like federated search systems, with details such as the individual's name and location fuelling queries to underlying social media, blogging and general Web searches. Disambiguation options are limited and there is often geographical confusion in the results, with the user being left to manually sift through the documents for the correct individual.

3. INFOSCOUT

Existing tools are insufficient because they do not adequately address the core problems faced by the investigator. Subject centric investigations rely heavily on disambiguating namesakes in order to acquire information about a particular individual, rather than people with the same name. This can be achieved with advanced query formulation, however this requires some degree of search expertise on the part of the investigator. Similarly, while existing solutions allow for faster and more convenient searching of multiple resources, the user must still manually inspect all documents, or their snippets, in order to identify relevant results, and extract the information, with minimal support for actually capturing evidence.

The requirements of the investigator can be described in simple economic terms: *they wish to maximise the recovery of information relating to the subject and the case, while minimising the amount of effort required.* InfoScout, the system described in this work, aims to address the investigator's difficulty with the large volumes of data found on the Web, allowing users to pinpoint information more readily, without the need for advanced query formulations or large quantities of existing knowledge.

InfoScout possesses the following features:

- Automatic page fetching and screenshot acquisition
- Content, entity and key term extraction
- User relevance feedback and re-ranking of search results
- Entity and term filtering of search results
- Case management

The core of InfoScout is designed around allowing the user to identify relevant documents quickly by re-ranking results based on user feedback. This allows the user to focus their time on extracting information from key documents, as opposed to locating such documents in the first place.

Initially, the user is presented with a simple interface which allows cases to be added, deleted, edited or continued. Each case has a subject as their focus, to which all search actions should pertain. The user enters ad-hoc queries and is then presented with a list of document screenshots, and links to said documents, for relevance judgements. This feedback

[2] https://www.copernic.com/en/products/agent/

[3] http://www.paterva.com/web6/products/maltego.php

[4] http://osintinsight.com/oryon.php

[5] https//pipl.com/

[6] http://www.spokeo.com/

is then used to build a ranking query which re-ranks the results using relevant terms and entities extracted from relevant documents. This inherently allows for person disambiguation as it allows for features of the users social network, location, hobbies and associated institutions to play a role in determining the order of documents. By marking a document with the entity *NASA* as relevant, it is more likely that further documents with this term are more highly ranked. Such re-ranking allows for results for *John Smith who works at NASA* to be distinguished from those of *John Smith who works for IBM*.

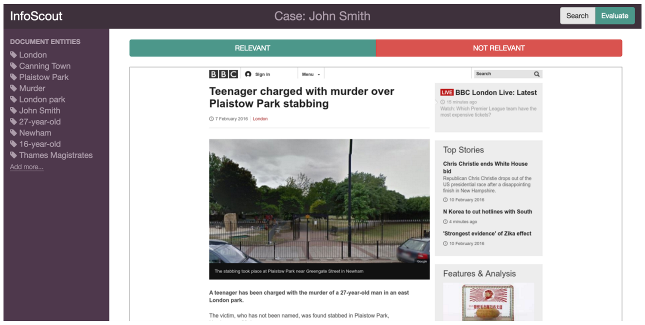
The interface for relevance feedback is simple, taking the form of a sequential suggestion of documents. The user is presented with a screenshot of the document and is given the option to mark the document as relevant, non-relevant, or to skip it altogether, see Figure 1. The re-ranking itself is transparent to the user.

The user must also be allowed to direct the search as they find relevant pieces of information about the subject. Adhoc queries can be submitted with new query terms, which will then be ranked using the information from previously judged documents. The user can choose to be presented with individual documents for judgement, or can explore the list of results as with a standard interface for Search Engine Results Pages (SERPs). Additionally, users may make use of extracted entities and terms in order to filter documents, as in Figure 2. This is useful in the case that the user wants to focus the investigation on a particular entities or keywords, such as the city or employer of the subject. Filters are applied on the standard SERP view.

Documents which are marked as relevant are also stored as part of the case, facilitating case management while also preserving the information as it appeared at the time of capture. This particular functionality mitigates the effort of manually tracking screenshots and documents in word processors, as with investigators in McKeown et al. [8] as well as mitigating later attempts at evidence obfuscation by the subject.

Search results are provided solely by the Bing search engine, as it allows for results from a wide range of Web results and includes publicly indexed social media pages. Early iterations of InfoScout used a federated approach, however this was discovered to be impractical, with API limitations and social media privacy settings negating the benefits. In practice, modern social media APIs are inherently privacy aware, such that less information is found programmatically than would otherwise be discovered by visiting the profile using a browser. Social media access in InfoScout leverages the Bing search engine, which may be tailored with appropriate queries[7]. Additionally, as was likely the case with Copernic Agent, the maintenance of a large number of APIs, which are likely to change over time, was considered an unnecessary overhead. Screenshots are provided to the user in order to facilitate rapid decision making of irrelevant documents while avoiding technical limitations of other means of directly displaying the document in the application, such as with iframes.

InfoScout facilitates the Open-Ended portion of the investigation, which is the most time consuming element of a case. In doing so, the cost of integrating existing, fixed, resources is mitigated, with the more constrained scope providing a cleaner and more focused solution. Similarly, no assumption is made about the financial position of the user or their subscriptions/access to popular industry resources, such as `192.com`.



Teenager charged with murder over Plaistow Park stabbing

7 February 2016 | London

The stabbing took place at Plaistow Park near Greengate Street in Newham

A teenager has been charged with the murder of a 27-year-old man in an east London park.

The victim, who has not been named, was found stabbed in Plaistow Park,

LIVE BBC London Live: Latest

15 minutes ago

Watch: Which Premier League team have the most expensive tickets?

Chris Christie ends White House bid

Republican Chris Christie drops out of the US presidential race after a disappointing finish in New Hampshire.

10 February 2016

N Korea to cut hotlines with South

4 minutes ago

'Strongest evidence' of Zika effect

10 February 2016

Figure 1: User document feedback, documents are suggested for judgement one at a time from the top of the ranking.

3.1 Implementation

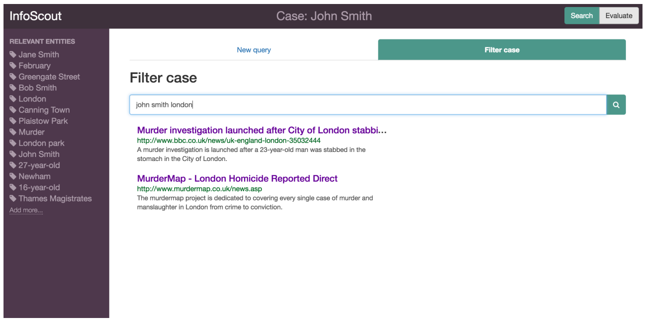
InfoScout is designed to be a Web application which can be run from any browser, with core components being hosted on the cloud as part of a distributed service. The core system is the OSINT API, which is both used to coordinate the various services and operations, as well as serve as the Web server which the user connects to. The API is implemented as a *Node JS* application, built using the *Express* framework. Re-ranking and filtering of Bing results is achieved by building an intermediate, local, index of Bing results, which makes use of the *ElasticSearch* platform. A consumer/provider pipeline in the OSINT API processes individual Bing result URLs using the *Alchemy Language* API[8], which allows the service to scale with large volumes of results. The parsed content of the documents, as well as their entities and most discriminative terms (within-document, ranked by their TFIDF) are stored in separate fields in the index. These fields facilitate later re-ranking and querying. Screenshots of the remote pages are captured using *Screenshotlayer*[9], which also stores the resulting images. Case information is stored using the *IBM's Compose MongoDB* service.

Document re-ranking is achieved by building ranking queries for submission to the ElasticSearch index. These ranking queries take the form of ElasticSearch Boosting Queries, which are comprised of a list of positive and negative document terms and entities. Modules within the OSINT API build this query in stages from the list of judged documents on the local index. The first stage aggregates scores for all terms and entities in the index for judged documents, with relative weights being calculated by within-document relevance, as determined by the Alchemy API. The second stage dynamically bins terms and entities by score, allowing for appropriate boosts to be applied to each bin. The final bins are then used to construct the re-ranking Boosting Query which is submitted to the ElasticSearch index for an updated document ranking. The result is that the most discriminative terms and entities from judged documents are used to weight

[7] Such as searching for John Smith's Google+ profile with the query prefix: site:plus.google.com

[8] http://www.alchemyapi.com/products/alchemylanguage

[9] https://screenshotlayer.com/



Murder investigation launched after City of London stabbi...
http://www.bbc.co.uk/news/uk-england-london-35032444
A murder investigation is launched after a 23-year-old man was stabbed in the stomach in the City of London.

MurderMap - London Homicide Reported Direct
http://www.murdermap.co.uk/news.asp
The murdermap project is dedicated to covering every single case of murder and manslaughter in London from crime to conviction.

Figure 2: Filtering of search results by applying ad-hoc filters allows the user to focus the search within retrieved documents.

new documents for the investigator, facilitating discovery of pertinent information. While both terms and entities are present in the re-ranking query, the boost weightings are biased towards entities as these are fundamental to this kind of search. Similarly, filtering of results is achieved by building Bool(ean) Queries for the ElasticSearch index, making use of the extracted entities and document content.

4. SUMMARY AND FUTURE WORK

We have presented InfoScout, which is intended to reduce the time spent searching public resources for subject-centric information. The interface is minimalistic and the inner workings are transparent to the user, mitigating requirements for training and search expertise. Users can direct their search while reducing the number of irrelevant documents that they encounter as they pursue a particular subject or critical piece of information.

While this system is essentially a proof of concept, with enough resources it would be possible to build a new entity centric index in the same manner as the Bing search engine. This would mitigate the need for a secondary index and mid-investigation page fetching, while providing more flexibility. Additionally, more advanced relevance feedback and re-ranking methods could be employed, such as by maintaining language models for relevant/irrelevant documents. However, this would require further research into domain specific re-ranking in order to determine what is most effective in Web based, entity centric, environment.

Additional features to be implemented will further improve the time saving potential of the application by providing more intelligence via extended automatic document processing. Information extraction techniques and clustering can be utilised to generate entity profiles and associated documents, allowing the user to further focus their search on a particular namesake. Entity co-citation and query expansion using known name variants would further bolster the person disambiguation process. Case management features will also be expanded to allow for entity and document relationship handling, with a graph based visualisation of the investigation.

While Bing is a powerful search engine, the user must still have some degree of acquaintance with its advanced operators in order to be maximally effective. To this end, it may be useful to facilitate advanced query building for the user by automatically populating site and boolean operators based on user input.

Finally, an evaluation of the system should be conducted in order to quantify its effectiveness. This could be broken into two parts: i) An evaluation of effectiveness of the re-ranking and user feedback, which would require the generation of an appropriate ground truth corpus; ii) A usability and task-based assessment conducted with industry professionals. The former would allow offline tweaking and optimisation of the underlying mechanics, while the latter would investigate the appropriateness of the interface design and determine the real world impact on the time/effort investment of the investigator.

5. REFERENCES

[1] E. J. Appel. *Internet Searches for Vetting, Investigations, and Open-source Intelligence*. Taylor and Francis Group, 2011.

[2] S. Attfield and A. Blandford. Improving the cost structure of sensemaking tasks: Analysing user concepts to inform information system design. In *Proc. of INTERACT '09*, page 532–545, 2009.

[3] Department of Homeland Security. DHS terrorist use of social networking facebook case study | public intelligence, 2010.

[4] J. Grasz. Number of Employers Passing on Applicants Due to Social Media Posts Continues to Rise, According to New CareerBuilder Survey - CareerBuilder, June 2014.

[5] R. Guha and A. Garg. Disambiguating people in search. In *Proc. of the 13th World Wide Web Conference,*, 2004.

[6] H. Joho, L. A. Azzopardi, and W. Vanderbauwhede. A survey of patent users: An analysis of tasks, behavior, search functionality and system requirements. In *Proc. of IIix '10*, page 13–24, 2010.

[7] Lexis Nexis. Social Media Use in Law Enforcement: Crime prevention and investigative activities continue to drive usage., Nov. 2014.

[8] S. McKeown, D. Maxwell, L. Azzopardi, and W. B. Glisson. Investigating people: A qualitative analysis of the search behaviours of open-source intelligence analysts. In *Proceedings of the 5th Information Interaction in Context Symposium*, IIiX '14, pages 175–184, New York, NY, USA, 2014. ACM.

[9] NATO. NATO open source intelligence handbook, 2001.

[10] A. Spink, B. J. Jansen, and J. Pedersen. Searching for people on web search engines. *Journal of Documentation*, 60(3):266–278, 2004.

[11] L. Šubelj, Š. Furlan, and M. Bajec. An expert system for detecting automobile insurance fraud using social network analysis. *Expert Syst. Appl.*, 38(1):1039–1052, Jan. 2011.