available at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/diin

Digital Investigation

Unsupervised discovery of relations for analysis of textual data

A.L. Louis[a,b,*], A.P. Engelbrecht[b]

[a] Council for Scientific and Industrial Research (CSIR), Meraka, Building 43, Meiring Naude Road, Pretoria 0001, South Africa
[b] Department of Computer Science, University of Pretoria, Lynnwood Road, Pretoria 0002, South Africa

ARTICLE INFO

Article history:
Received 22 August 2009
Received in revised form
15 August 2010
Accepted 30 August 2010

Keywords:
Evidence discovery
Text analysis
Text mining
Information extraction
Relation discovery

ABSTRACT

This paper addresses the problem of analysing textual data for evidence discovery. A novel framework in which to perform evidence discovery is proposed in order to reduce the quantity of data to be analysed, aid the analysts' exploration of the data and enhance the intelligibility of the presentation of the data. The framework combines information extraction techniques with visual exploration techniques to provide a novel approach to performing evidence discovery, in the form of an evidence discovery system. By utilising unrestricted, unsupervised information extraction techniques, the investigator does not require input queries or keywords for searching, thus enabling the investigator to analyse portions of the data that may not have been identified by keyword searches.

A preliminary study was performed to assess the usefulness of a text mining approach to evidence discovery from a text corpus in comparison with a traditional information retrieval approach. It was concluded that the novel approach to text analysis for evidence discovery presented in this paper is a viable and promising approach for consideration in digital forensics. The preliminary experiment showed that the results obtained from the evidence discovery system are sensible and useful.

1. Introduction

The use of the many forms of computer and communication devices today results in the generation of enormous pools of data which, in turn, enables criminal investigators to make use of this data to uncover evidence. A great deal of digital data is linguistic in nature (e.g. human languages, programming languages, and system and application logging conventions) (Beebe et al., 2007). The expressivity of language makes these files rich information sources, which makes textual data and consequently textual evidence very important in digital investigations. The obvious and immediate difficulty which confronts any person wishing to make use of the potential which digital evidence holds, is that analysing and processing the ever-growing volumes of data, and linking the unstructured data together for use as evidence in a trial involves a tremendous amount of work. Examining and organising all of the data which digital evidence yields is a major challenge facing all investigators. For electronic evidence to become properly useful to investigators, accurate, efficient and rapid means of data analysis need to be developed (Chen et al., 2004).

It is hypothesized in this paper that information extraction techniques combined with visual exploration techniques can assist in identifying suspects and events, and the relations between these entities which could assist an investigator to: piece together the story surrounding a crime, create hypotheses

[*] *Corresponding author*. Council for Scientific and Industrial Research (CSIR), Meraka, Building 43, Meiring Naude Road, Pretoria 0001, South Africa

E-mail addresses: alouis@csir.co.za (A.L. Louis), engel@cs.up.ac.za (A.P. Engelbrecht).

1742-2876/$ – see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.diin.2010.08.004

or potential leads for further investigation, or identify pieces of data which could form useful supporting evidence in a trial. This research investigates the adaptation and application of text mining research to evidence discovery to address the need for a tool to aid analysts to more quickly, efficiently and accurately analyse data to reveal truly useful information. It is believed that the proposed framework provides potential as a tool in digital forensic investigations.

The primary contributions of this work are:

- a novel framework in which to perform evidence discovery,
- a statistically based model for performing relation extraction (model A), and
- an alternative linguistically based model for performing relation extraction (model B) as components for the evidence discovery framework.

A preliminary study was devised to assess the usefulness of a text mining approach to evidence discovery in comparison to traditional information retrieval approach, utilising Agatha Christie's novel entitled 'The Mysterious Affair at Styles' as a dataset. The preliminary experiment showed that the approach advocated in this paper can therefore be successfully applied to the analysis of textual data for digital forensics.

2. Background

Computer software exists that can assist in many phases of a digital investigation. However, it is important to note that this software is not designed to solve crime. The primary goal of these systems is to reduce investigation time and complexity (Abraham, 2006).

Expression based search methods currently dominate the software tools used for the analysis of digital data. These tools focus on providing advanced search technologies with which an investigator can retrieve potential evidence, thus search technology is called information retrieval. An investigator will use his/her experience and background knowledge of the case to select appropriate search terms and criteria to find clues in the data as to which data should be investigated in more detail, or to find data which will offer suggestions as to how to proceed with the investigation. However, it is not uncommon in digital investigations for very little to be known about the case or the collected data prior to analysis.

Information retrieval (IR) is only as good as the query terms used, which means that the information retrieved in an investigation is limited to the background knowledge of the case and extended search terms from the investigator's personal experience. Arguably the dependence on query terms and their many combinations makes searching for digital evidence very time consuming and inefficient.

A move away from information retrieval techniques includes text summarisation, document classification or clustering and text mining (Hotho et al., 2005; Dozier and Jackson, 2005; Fan et al., 2006; de Waal et al., 2008). In a perfect world, a forensic system could "discover" data that seems suspicious. A discovery system which does not rely on user input for query terms could therefore assist in speeding up the analysis phase of the investigation. This has led researchers to investigate different approaches to finding evidence.

Promising results have been achieved by text mining in the biosciences application area. Perhaps the most cited example is Don Swanson's work on hypothesizing causes of rare diseases by looking for indirect links in different subsets of the bioscience literature (Swanson, 1987, 1991).

Dr. Liebman is convinced that new cures could someday emerge for breast cancer if only someone could read all the literature and synthesize it (Guernsey, 2003). Thus, Dr. Liebman enlisted the use of text mining software to 'read' medical journal articles. The preparation for Dr. Liebman's project took months to build a framework of knowledge required as input to the system, on the subject of interest, namely breast cancer (Guernsey, 2003). The output of the software is a visual map of extracted concepts, which lead Dr. Liebman and his team down new pathways, which they could then test scientifically.

The promising results of text mining in the biosciences field provides the motivation to explore the application of text mining methods as an alternative approach to finding evidence in text data.

3. A framework for evidence discovery

One of the primary difficulties with finding relevant information quickly using the current tools and methods is that they tend to be query- or search- term driven and these terms, by definition, rely on the analyst having prior or background knowledge about the data to be analysed. Accordingly, search term analysis is necessarily restricted to the portions of data relating to the query terms. No information is obtained about data which does not fit the query terms. To make the investigation process more efficient, tools need to be developed that will reduce the quantity of data to be analysed, aid the analysts' exploration of the data and enhance the intelligibility of the presentation of the data.

Text mining techniques enable text data analysts to explore text data, which enables them to structure, find, or extract the information they require more quickly and efficiently. At present, automatic information extraction systems require large amounts of domain knowledge to be embedded into the system to make the system usable for a new domain (Sekine, 2006). The large quantity of domain knowledge required makes the portability of text mining systems to different domains extremely limited thus holding back the usefulness of text mining and, in particular, information extraction systems. Surdeanu and Harabagiu (2002) identified that domain independent information extraction systems using un-annotated text still face many challenges. This is because the performance of current language analysis tools remains poor, causing system designs to require significant user interaction and making the design of a new system a time-intensive task (Surdeanu and Harabagiu, 2002).

This paper proposes a novel framework in which to perform evidence discovery in an attempt to meet these challenges. The framework combines existing information extraction and visual data exploration techniques to create a text graph of the most important concepts and associations extracted from the text. The text graphs provide an overview and general representation of the text, which may help the investigator explore and analyse the text more efficiently. An overview of the framework is first described, followed by descriptions of each component of the framework namely, document pre-processing, relation discovery, and text-graph creation and visualisation.

It is not uncommon in digital investigations for very little to be known about the case or the collected data prior to analysis. A discovery system which does not rely on user input for query terms could therefore assist in speeding up the analysis phase of the investigation.

Due to the sensitive nature of case datasets, it is difficult, especially in the South African context, to obtain authentic data for research purposes. de Waal et al. (2008) performed a case study on an authentic forensic dataset, however the usefulness of their technique could not be fully evaluated as the investigators who provided the dataset had not themselves generated usable evidence from the data. For these reasons a fictitious dataset was used in this research to establish the feasibility of the framework and subsequent evidence discovery system.

The murder mystery novel by Agatha Christie entitled The Mysterious Affair at Styles was chosen as a dataset as this novel is available for free download in electronic plain text format from Project Gutenberg (Lebert, 2008). An Agatha Christie novel was chosen because it was felt that it represents the complex wealth of knowledge that is present in real evidentiary data. Agatha Christie's novels are characterised by complex plots involving a crime and several characters, all or many of whom are typically presented as potential suspects in the reader's mind. Christie's mysteries are achieved by reserving the identity of the true criminal or criminals until the conclusion of the novel. Because the stories are complex and so many characters are potential suspects, there is an element of noise surrounding the main thread of the crime. Although a literary work is not written in the same style as business related documents, letters and memos, the datasets used to develop linguistic analysers, parsers and information extraction algorithms generally comprise a very large collection of short news articles. Using a literary work will therefore test the domain independence of the existing NLP tools, which will give an indication of their potential performance on a real criminal dataset.

Like a novel, a true-life criminal act and the events and suspects surrounding it form a story, or narrative, consisting of actors, actions and relations. Fig. 1 shows the parallels between a crime and a story.

Visualising extracted concepts and/or named entity (NE)s and their relations in a text graph, could assist an investigator to piece together the story surrounding the crime. Thus, relation extraction techniques seem apt for the challenge. By removing the reliance on an input query required by traditional information extraction (IE) systems, the evidence discovery system enables the investigator to explore portions of the data that they would not be able to using traditional search-based methods.

The Evidence Discovery system proposed in this dissertation consists of a few pre-processing steps followed by relation extraction, which is then used to create text-graphs of the story. Fig. 2 shows the components of the framework and the flow of data.

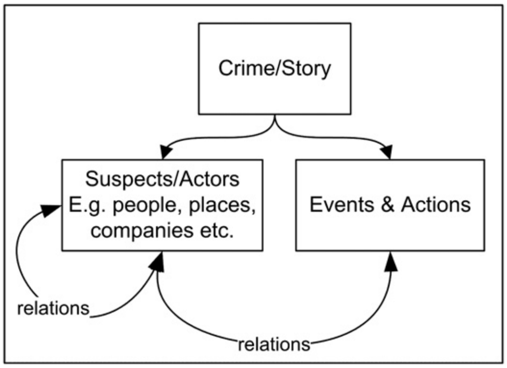


Fig. 1 – Parallels between a crime and a story.

The following sections describe each component of the framework illustrated in Fig. 2. First the text documents need to be processed with some basic text analysis. Relation extraction techniques are then employed to extract concepts from the text and their relations or 'associations' to each other. A text graph is then created from the extracted concepts and relations, which is then visualised and presented to the user in an interactive interface, which enables the user to explore and evaluate the associations in the graph.

3.1. *Document pre-processing*

Computers handle text as simple sequences of character strings. Therefore pre-processing algorithms are used to extract and store the information from text documents in a data structure that is more appropriate for further processing than a plain text file (Hotho et al., 2005).

Text pre-processing algorithms are used to convert text from a simple sequence of character strings into tokens and chunks so as to identify syntactic elements and to store the data in a form that is more appropriate for further processing than a plain text file. Different types of pre-processing can be done, and the appropriate pre-processing techniques should be chosen depending on the text analysis algorithm which is intended to be used.[1]

Linguistic pre-processing incorporates a number of methods and algorithms, most of which originated in the field of natural language processing (Manning and Schütze, 2001; Jurafsky and Martin, 2008). These methods extract or label additional information about the words or text to reveal information about the syntactic structure, such as:

- Part of speech tagging determines the part-of-speech of each word in a sentence based on its definition and context, and applies a label or tag (e.g. noun, verb, adjective, etc.) to each word. While there are a number of both stochastic and rule based approaches, the most well-known algorithm is the Eric Brill tagger (Brill, 1992).

[1] Manning and Schütze (2001) provide a comprehensive introduction to statistical natural language processing (NLP).

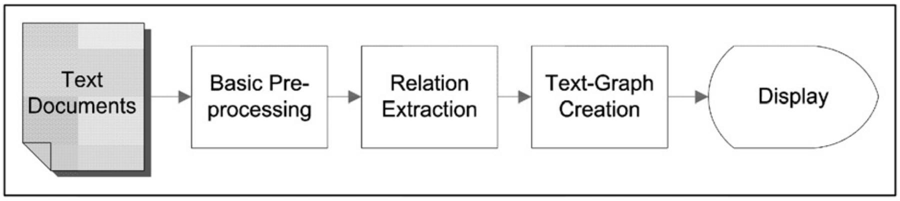


Fig. 2 – Parallels between a crime and a story.

- Phrase recognition, known as chunking, segments sentences into their subcomponents by grouping or 'chunking' adjacent words into phrases. For example, the chunks "the tile rooftop" and "especially slowly" would be recognised as noun and adverbial phrases respectively.
- Word sense disambiguation, which is the task of discerning the meaning of a word in a given context to resolve the ambiguity of which sense of the word is intended. For example, 'bank' in the noun form may refer to the 'financial institution' or the 'border of a river', or in the verb form it may mean 'to cash in', or 'to tilt'.
- Parsing, which assigns the relation of each word in a sentence to each other word and represents the sentence in a tree structure. Each word is annotated with its grammatical type and its function in the sentence, e.g. subject, object, etc. (Jurafsky and Martin, 2008).

The relation extraction algorithms used in this paper require two forms of text pre-processing. First the text documents are processed by a part of speech tagger, using Infogistics' NLProcessor software (Infogistics, 2001). NLProcessor first tokenises the text, segmenting the stream of characters into separate parts or tokens to obtain the basic units of paragraphs, sentences and words. NLProcessor then applies an algorithm to determine the part-of-speech of each word in a sentence based on its definition and context and applies a label or tag (e.g. noun, verb, adjective, etc.) for each word, based on the Modified Penn Treebank Tag-Set. Fig. 3a shows an example of a sentence that has been tagged with part of speech tags. WordNet (Fellbaum, 1998), an electronic lexical database, is then used to tag each word with its lemma, based on its part of speech to represent each word in its normalised form.

3.2. *Relation discovery*

After the text has been pre-processed and marked up with syntactic tags, these tags can be used to perform further text analysis. Relation discovery aims to find and extract concepts and their relations or associations.

This paper presents two relation discovery models. The first (Model A) relies on statistical measures of association and co-occurrence, whereas the second (Model B) aims to exploit the syntactic structure and linguistic characteristics of the text, such as named entity recognition (NER) and full syntactic parsing. These two relation discovery models are discussed in detail in Sections 3.2.1 and 3.2.2.

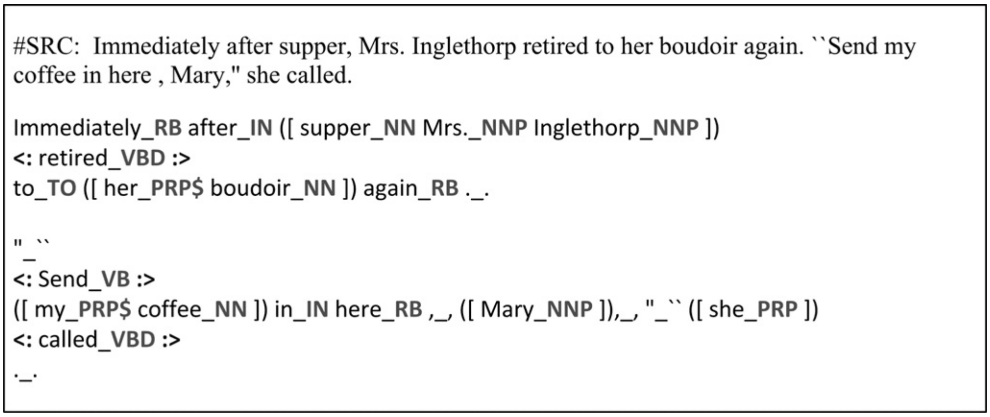
Normally, extracted information would be stored in tabular form, where events or similar information would be grouped together in tables. The relevant table can then be retrieved when desired, using a query. For example, a query with the keywords "merge, merger, acquisition, purchase, buy" may produce a table in the format shown in Table 1.

In this research the extracted concepts and relations are stored in the form of text graphs to explore the data, rather than to retrieve it. The text-graphs can then be visualised and explored.

3.2.1. *Model A*

A concept is an "idea" or "unit of thought", and is said to be complex if it consists of more than one word. A complex concept is formed from co-occurring words that are strongly associated with each other, as presented and described in Louis (2009).

This model takes a statistical approach to discovering relations based on co-occurrences of concepts. Model A utilises a 'bag of words' approach based on the assumption that



```
#SRC: Immediately after supper, Mrs. Inglethorp retired to her boudoir again. ``Send my
coffee in here , Mary," she called.

Immediately_RB after_IN ([ supper_NN Mrs._NNP Inglethorp_NNP ])
<: retired_VBD :>
to_TO ([ her_PRP$ boudoir_NN ]) again_RB ._.

"_``
<: Send_VB :>
([ my_PRP$ coffee_NN ]) in_IN here_RB ,_, ([ Mary_NNP ]),_, "_`` ([ she_PRP ])
<: called_VBD :>
._.
```

Figure 3a – Example of sentence pre-processing.

Please cite this article in press as: Louis AL, Engelbrecht AP, Unsupervised discovery of relations for analysis of textual data, Digital Investigation (2010), doi:10.1016/j.diin.2010.08.004

Table 1 – Example of a typical table retrieved from an IE query.

Company	Money	Date
ABC Bank	About $1.6 billion	
BB Corp, XYZ inc.	$3 million	May
JJ Holdings	$286 million	
World Finance Corp.	About $400 million	7 July

a relation occurs between two concepts if these concepts co-occur within a certain boundary or window with a frequency above a certain threshold. Based on the idea of market basket analysis whereby associations are found between purchased products, model A aims to find 'association links' or relations between the frequently co-occurring concepts in the text.

A relation is said to exist between two concepts if the concepts co-occur within a certain window size or text segment having a support of a chosen minimum number of occurrences. To form relations that make semantic sense, concepts were filtered so as to only include concepts tagged as nouns and adjectives.

The size of the text segment used can be varied in order to optimise the results of the text graphs. For the novel used in this work, segments consisting of one sentence were implemented, as this causes a strong association between the concepts. Increasing the size of the text segment will increase the distance between concepts and lessen the degree of association enforced. A sample of the relations extracted is shown in Table 2.

Examination of the associations shows that these are indeed sensible. \Afternoon and \will occur together five times and in each case the question of a will being drafted on the afternoon in question is discussed. The association between "coffee" and "strychnine" is intriguing and is summed up by one of the sentences in the novel, which states: "The present contention is that Mrs. Inglethorp died of strychnine poisoning, presumably administered in her coffee." Exploration of "Cynthia" and "Lawrence" merely shows that they are acquaintances, however in some cases a discovery of that nature could be the key to the mystery or crime.

3.2.2. *Model B*

Model B utilises a linguistic approach and aims to extract relations using NE extraction and IE patterns. To make use of the syntactic structure to extract relations some additional linguistic pre-processing steps, namely full syntactic parsing and NE extraction, are required. After the basic pre-processing, whereby the text is split into sentences and words, shallow parsing is performed using the Charniak parser (Charniak, 2005) which performs part of speech (POS) tagging and phrase chunking. The parsed text is then subjected to NE extraction and full syntactic parsing. In this dissertation, NE extraction was performed using the Java Extraction Toolkit (JET) (Grishman, 2007). The extracted NEs are put aside for later use in the relation extraction process.

In order to extract relations using an extraction pattern model, a dependency analysis, in the form of a full syntactic parse, needs to be performed. For the purpose of this analysis a sentence is represented as a set of asymmetric binary links between a word and its modifiers in the form of a dependency tree annotated with its grammatical type and its function in the sentence, e.g. subject, object, etc. This study uses a syntactic parser developed at New York University, the Grammatical and Logical Argument Framework (GLARF) (Meyers et al., 2001), to perform full syntactic parsing. GLARF is a typed feature structure framework for representing regularizations of parse trees. GLARF produces dependency trees with both the surface and logical relations. Where sentences are written in the active voice the surface and logical relations are the same. However in the passive voice the true relations are represented in the logical form. For example, in the sentence "The apple was eaten by John," the surface subject is "the apple" and the logical subject is "John." Therefore the logical relations are used for relation discovery. Fig. 3b shows the dependency tree for the sample sentence "Immediately after supper Mrs. Inglethorp retired to her boudoir again."

The extracted NEs in conjunction with the dependency trees are then used to extract relations using an IE pattern model. A number of extraction pattern models exist, each of which incorporates different amounts of information about the text. The extraction pattern model used determines the number of patterns that can be extracted.

The linked chains extraction pattern model was chosen for the evidence discovery system because it offered the best compromise between capturing too many patterns or too little information. The linked chains model Greenwood et al. (2005) defines a pattern as a pair of chains which share the same verb, but no direct descendants. The linked chains model

Table 2 – Sample of the extracted relations (Model A).

Concept 1	Associated with	Concept 2
Afternoon	↔	Will
Coffee	↔	Strychnine
Cynthia	↔	Lawrence
Cynthia	↔	Moment
Death	↔	Strychnine
Death	↔	Mother
Favour	↔	Will
Mrs. Inglethorp	↔	Strychnine
one	↔	Will

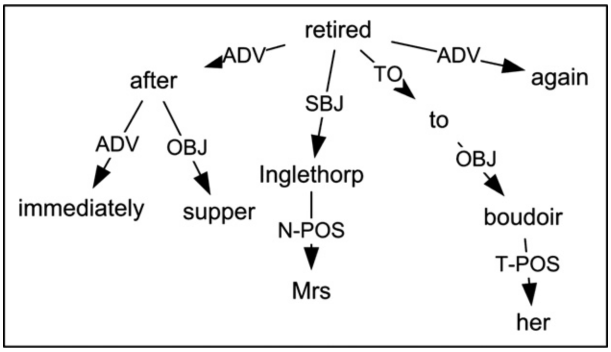


Figure 3b – An example of a dependency tree.

captures both the relationships beyond clausal boundaries as well as the link between arguments or subjects and objects of the verb.[2]

From the dependency tree in Fig. 3b, a total of 29 linked chains can be extracted. For the purposes of relation extraction for the evidence discovery system, relations of the form \subject-relation-object" were chosen to ensure that relations contain concepts rather than other parts of speech such as adjectives or adverbs. Thus only patterns that contain both a subject and an object are extracted. This results in only two linked chains from the dependency tree in Fig. 3b:

- (PRED: retired (SBJ: Inglethorp (N-POS: Mrs))(ADV: after (OBJ: supper)))
- (PRED: retired (SBJ: Inglethorp (N-POS: MRs))(TO: to (OBJ: boudoir (T-POS: her))))

However, it was found that the syntactic parser could not always parse the entire sentence and would leave fragments, some of which did not contain a verb, and therefore no patterns could be extracted from those fragments.

The extracted NEs were then inserted back into the extracted relations as tags to indicate their identity and class. For each entity, the JET records the type of the entity (PER, ORG, GPE, LOC, FAC, VEH, WEA), subtype, class, and all the textual mentions of that entity. The complete set of annotations for English and their explanations can be found in the ACE (Automatic Content Extraction) English Annotation Guidelines for Entities LDC (2005). The co-referencing of NEs is a difficult task for a novel or literary work with many characters, especially where names are shared and are often referred to using pronouns. It was found that the number of errors in the co-referencing of NEs were more harmful than helpful. Therefore, it was decided that the NEs would be tagged, but not co-referenced.

At least one relation can be extracted for each sentence in the text, which results in too many relations for the user to be able to grasp. Thus weights were applied to each subject and object of the extracted relations using the term frequency-inverse document frequency (tf.idf) weight (Salton and McGill, 1983). The tf.idf is a statistical measure used to evaluate how important a word or term is to a document in a collection or corpus. Only relations whose subject and object are ranked as important were retained, reducing the number of relations to a small set of important relations, which is more manageable for the user. The document collection was assembled from 80 books downloaded from the Project Gutenberg (Lebert, 2008) repository.

Fig. 4 shows the weights of all of the words in the text in descending order. The higher the weight the more important the word is to the document. In order to determine which words are significant and should be retained, a threshold value for the tf.idf was chosen. The threshold was chosen to eliminate 98% of the values by calculating the mean and standard deviation of the weights and setting the threshold to one standard deviation above the mean. Fig. 5 shows these values. The two highest weights, assigned to the two main character's names, were found to be extreme values because they occurred very frequently in the relevant book and did not occur in any of the other books in the collection. For that reason those two weights were excluded from the calculation of the threshold. Thus the threshold was set to a value of 1.0856E^{-4}. Terms that were identified as NEs were assigned a weight of 1.0 to ensure that they are not filtered out. For example, the name "John" may have a very high document frequency causing a very low tf.idf. However, NEs are assumed to have a high evidentiary value to investigators and therefore they are assigned a very high weight. Only relations whose subject and objects were weighted above the threshold were then retained.

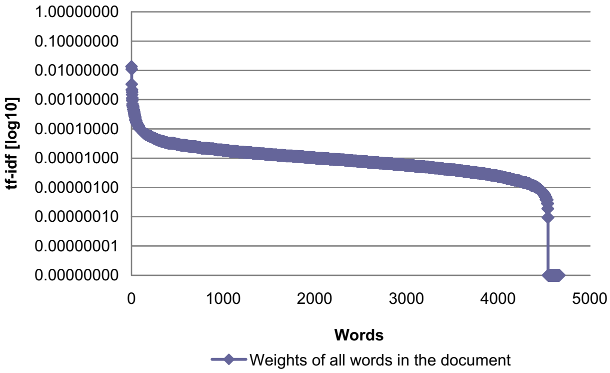


Fig. 4 – Tf.idf weights for all words in the document.

Table 3 shows a sample of the extracted relations from the novel used in this research. The resultant relations, after filtering, are shown to be sensible relations between the subjects and objects, despite the removal of words from the original sentence in the extraction process. Although some errors occurred in the NE tagging, the relations involving NEs give the reader information about the NEs and can describe the associations between NEs. Relations that do not include any NEs are shown to be interesting and add detail to the story behind the data.

Text-graphs were then created to illustrate the ties between the extracted relations based on common NEs, subjects, and objects. The subjects and objects of the relations are represented by nodes, with labelled directed links representing the action of the relation using an arrow from the subject to the object. Multiple relations between nodes may lead to more than one link between two nodes. Fig. 6 shows an

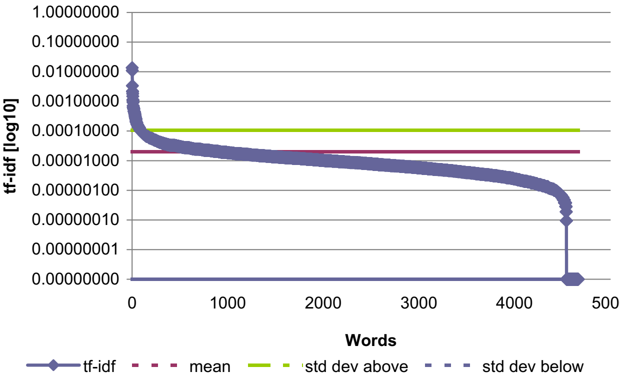


Fig. 5 – Tf.idf threshold calculated for the document.

[2] The S-expression form of the linked chains model is noted as (PRED: - (ARG1: - ((ARGn: -))) (ARG2: -((ARGm: -)))).

Table 3 – Sample of the extracted relations with NE tagging after filtering (Model B).

Subject	Relation	Object
Cavendish (PER)	in mother-in-law's	Room (PER)
I	tell	You
coco	contained no	Strychnine
John (PER)	views concerning	Bauerstein
She	Said to	John (PER)
Japp	accompanied	Car (VEH)
I	descended from the	Train (VEH)
Cynthia (PER)	protegee of the mother	Daughter (PER)

example of a text-graph created from a selection of extracted relations from the text. This example shows that the nodes 'Inglethorp' and 'John' are both of type person (PER) and are both related to the object 'room'. Their relations with 'room' are 'had in the' and 'opened door of' respectively.

4. Text graph creation and visualisation

Visualisation is a very important aspect of any system that is intended to be used by a human, as humans are very receptive to interpreting graphical information. To provide a means for the user to explore and evaluate the extracted relations, text graphs are created from the concepts and their relations. These text graphs are then visualised and presented to the user in an interactive interface.

The text graphs are created in the form of a highly connected graph where nodes represent concepts and links are drawn to show associations between concepts. The graph is then displayed on a graphical user interface (GUI), using a force-based algorithm based on the algorithms of Eades (1984) and Churcher et al. (2004), as presented and described in Louis (2009). Fig. 7 illustrates resultant text graphs created from the entire novel used for this example.

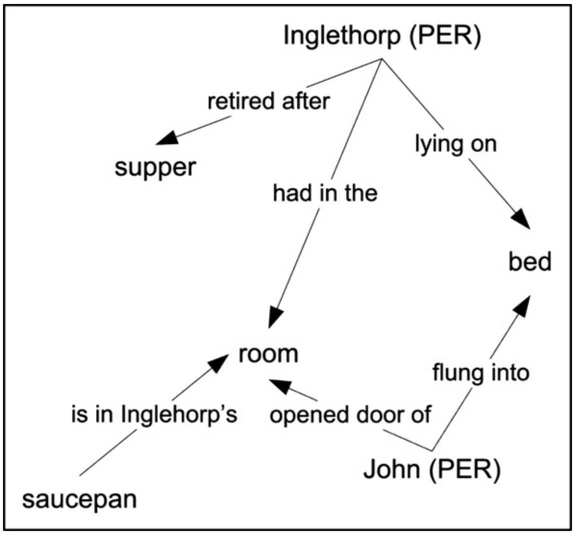


Fig. 6 – Example of a text-graph created from extracted relations.

'Key Concepts' are defined to be those concepts that are important to the story based on the fact that they are highly connected. Nodes that have more than a specified threshold, n_k, links are high-lighted as 'key concepts' to enable the user to see at a glance which concepts are important to the story in the novel. The number of relations or links that a node has can be used as an indicator of that node's importance within the text. The higher the number of links, the more important the node is. A histogram of the number of links for the nodes extracted from the novel in this example, for each of the relation extraction methods, is presented in Fig. 8. The histogram shows that the greater the number of links, the fewer the number of nodes having that many links. There is a rapid fall-off of the number of nodes as the number of links increases, which aids in choosing the threshold, n_k. For the purposes of this case study a threshold of $n_k \geq 4$ was chosen through trial and error, and can be optimized for corpora of varying sizes depending on the users' preference.

To understand the graph better, the user is able to interact with it in two ways. Firstly, where there are an immense number of nodes, the user may rearrange the nodes to organize, cluster, separate or spread out certain nodes in the graph. This is done by clicking on a particular node and dragging it to the desired position. Secondly the user may explore the context of the links. When a single node is selected a side panel on the GUI shows all of the sentences that created the association between that node and all the concepts that the node is connected to. If two nodes are selected, only the contexts of those links between those two nodes are shown. This helps the user to understand the meanings of the concepts as well as their association with each other. Fig. 9 demonstrates this functionality. By traversing the links, an investigator may explore relations between concepts that are connected through other concepts. The degree of separation is measured by the number of links between the two concepts. Thus threads of the story can be explored.

Model B of the relation discovery process identifies entities and NEs. If entities have been identified and labelled in the graph the GUI will display a list of these entities in the side panel. The user may then select an entity from the list to display a subgraph containing the selected entity and its links. In this way a user may examine entities and their links more closely. Fig. 10 shows the selection of the entity "Poirot (PER)" from the list and the corresponding subgraph.

5. Experimental design

An approach is needed to assess whether the product of the framework and models is useful and whether the graphical presentation of the data are useful to a user. If neither of the models produce results that are useful to a user, then it must be concluded either (a) that the novel approach to text analysis for evidence discovery presented here is not a successful/viable one or (b) that the implementation of that approach was not successful. If either of the models does produce results that are useful, then it can be argued that the approach advocated here can be successfully applied to the analysis of textual data for digital forensics.

Please cite this article in press as: Louis AL, Engelbrecht AP, Unsupervised discovery of relations for analysis of textual data, Digital Investigation (2010), doi:10.1016/j.diin.2010.08.004

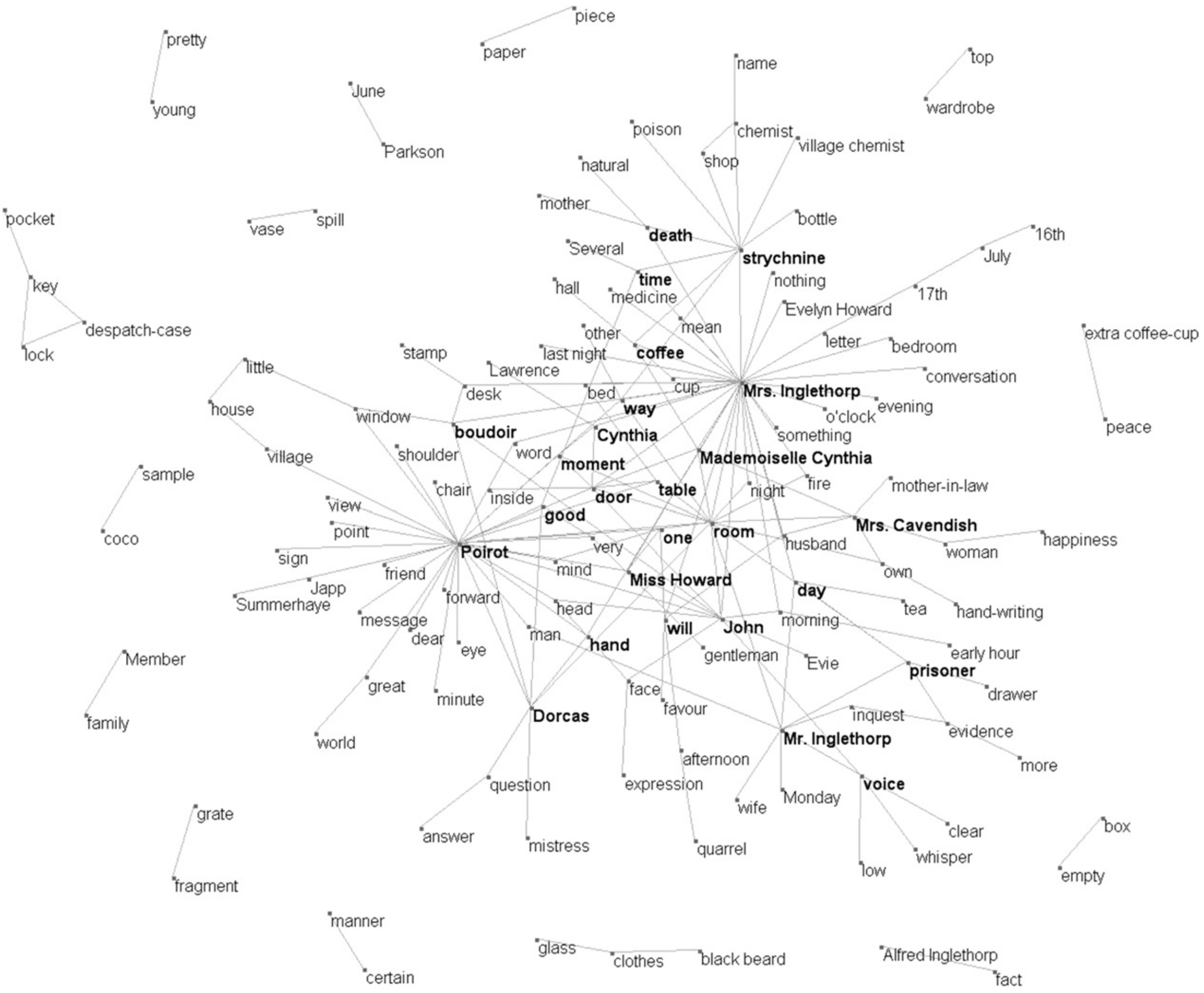


Fig. 7 – Visualisation of the Text Graphs created from the novel The Mysterious Affair at Styles.

Because a full-scale quantitative assessment of the performance of the models is beyond the scope of this study, a preliminary study was devised with the aim of determining whether the evidence discovery system produces sensible, useful and useable results. A survey of the literature showed that digital forensic analysis is predominantly performed using text string searching (Beebe et al., 2007). Text based searching therefore provides a useful comparator or control against which to compare the two models advocated here. The experiment is presented and discussed below and it is demonstrated that Model A and Model B are both equally as useful in this pilot study as search based analysis.

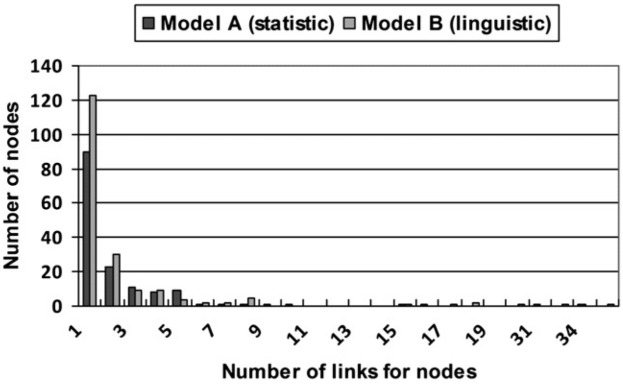
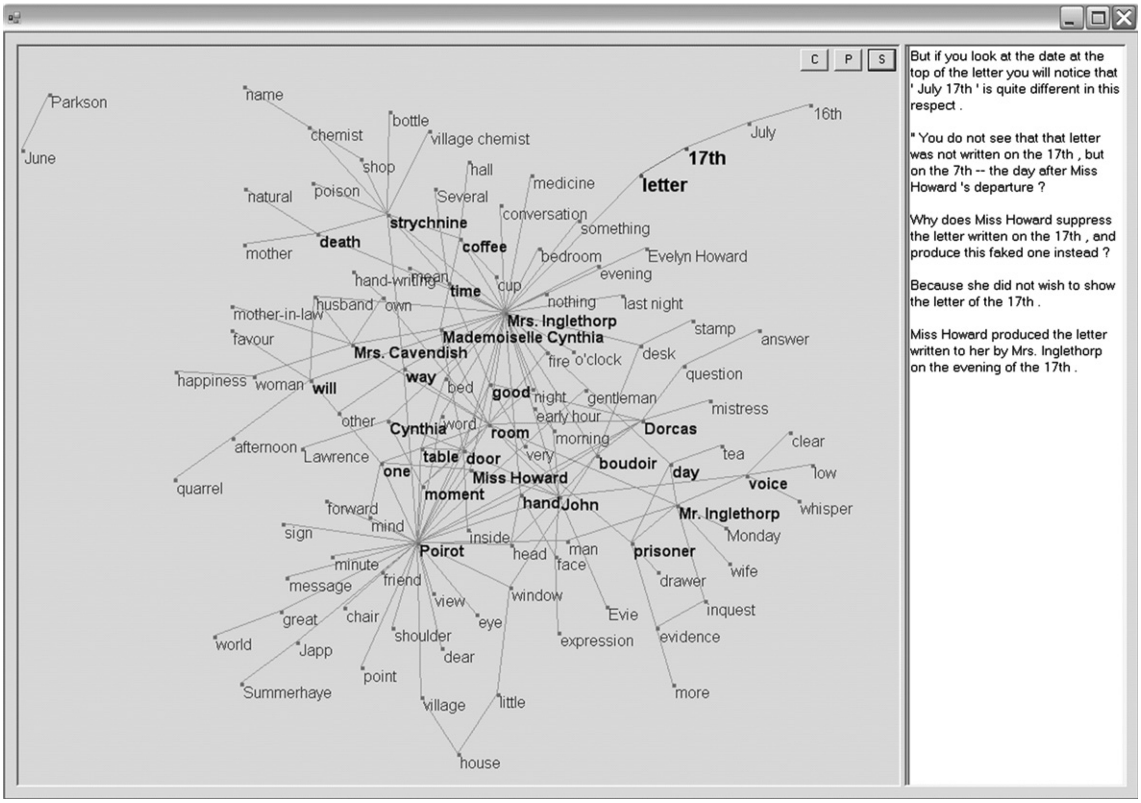


Fig. 8 – Histogram of number of links for the nodes extracted from the example.

One of the fundamental differences between IR and discovery is the requirement for initial background information. While a discovery system has the advantage of not requiring any initial insight into the data or input query terms, when evaluating the two systems against each other the users of the search method will require some initial background information to tell the user what to look for and to assist in choosing initial query terms.

In order to evaluate the analysis of data using each method, a user based test was designed based on the idea of a 'reading comprehension', whereby the user's knowledge and understanding of the text is tested. While a multiple choice question testing mechanism would provide a way to quantitatively evaluate the models, the main aim of this study is to investigate the effectiveness and usability of the evidence discovery system for evidence discovery. Therefore, the 'essay style' testing mechanism was chosen, because it simulates how the tool would be used in a real scenario. Thus,



But if you look at the date at the top of the letter you will notice that ' July 17th ' is quite different in this respect .

" You do not see that that letter was not written on the 17th , but on the 7th -- the day after Miss Howard 's departure ?

Why does Miss Howard suppress the letter written on the 17th , and produce this faked one instead ?

Because she did not wish to show the letter of the 17th .

Miss Howard produced the letter written to her by Mrs. Inglethorp on the evening of the 17th .

Fig. 9 – Side panel shows context of links between “letter” and “17th”.

a qualitative analysis will need to be done, to evaluate the performance of the evidence discovery system.

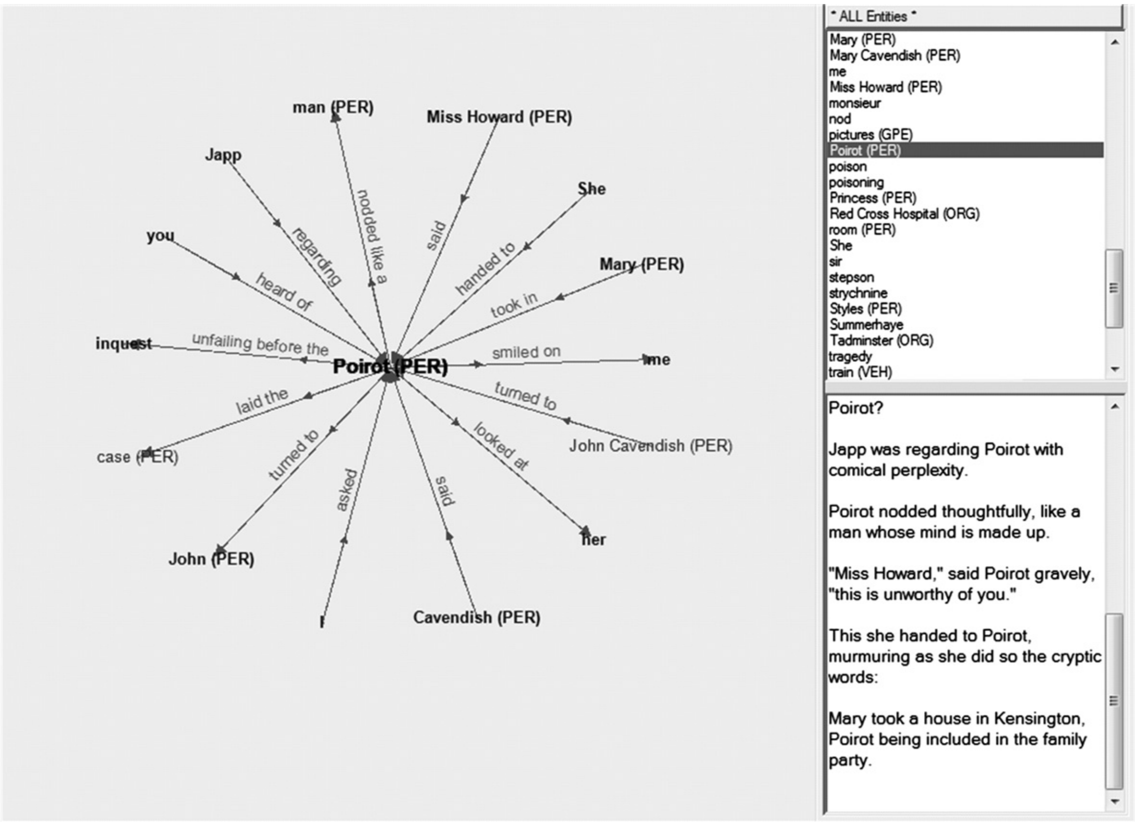
Quantitative methods typically depend on large samples in order to generalize with confidence from the sample to the population that it represents. However, according to Patton (2002), qualitative inquiry typically focuses on relatively small samples, even single cases, to permit enquiry into and understanding of an experience in depth. While one cannot generalise from single cases or small samples, one can learn a great deal from them (Patton, 2002). Given the qualitative nature of this study, a small samples size of four per group was chosen. This allows for each of the participants results to be analysed in depth to gain insight into how the user interacts with the system (Patton, 2002).

Twelve random participants, each of which had some prior experience of search-based analysis methods, were randomly divided into three groups of four. Each group was then randomly assigned to an analysis method:

- Group 1 tested the evidence discovery system with the statistically based relation discovery model, Model A. Within the experiment this method as a whole will be referred to as Model A.
- Group 2 tested the evidence discovery system with the linguistically based relation discovery model, Model B. Within the experiment this method as a whole will be referred to as Model B.
- Group 3 tested a search method: TextPad (HeliosSoftware Solutions, 2000) provides a search engine using UNIX-style regular expressions, with which the users could perform within document content searching.

The novel entitled ‘The Mysterious Affair at Styles’ by Agatha Christie was used as a dataset for this research. The novel presents a mystery concerning a murder. The nature of Christie’s writing is such that there are a number of red-herrings to confuse the reader, which present a challenge to the reader to solve the mystery of the crime. However, Christie typically reveals the manner in which the murder was committed and the identity of the murderer in a final paragraph or paragraphs. In order to make the novel more akin to a true dataset, this portion of the text was removed for all users.

Real evidence found in real data, in contrast to the fictitious ‘evidence’, can be unclear and uncertain and has a significantly greater noise component due to the data which is not regarded as having evidential value. For instance, only one small piece of evidence may be found in an entire hard drive, whereas the novel contains all of the evidence and facts required to solve the crime. It is therefore advantageous to use fictitious data for a comparative evaluation, because the evidence presented in the novel is clear and can be easily evaluated by reading the story. To assist the Group 3 users to choose their initial query terms, all of the users are told that the text is a novel about a murder mystery. None of the participants had read an Agatha Christie novel before the testing; in particular none had read ‘The Mysterious Affair at Styles’.



Poirot?

Japp was regarding Poirot with comical perplexity.

Poirot nodded thoughtfully, like a man whose mind is made up.

"Miss Howard," said Poirot gravely, "this is unworthy of you."

This she handed to Poirot, murmuring as she did so the cryptic words:

Mary took a house in Kensington, Poirot being included in the family party.

Fig. 10 – Selection of entity "Poirot" and its corresponding subgraph.

Before commencing the test a demonstration of the software was given to the participants, using the book of 'Peter Pan' by James M. Barrie, after which participants were given 5 min to become acquainted with the tools themselves, using the story of Peter Pan.

On commencement of the test, each user was allowed 1 h to analyse and explore the test dataset to find 'leads' and 'evidence' regarding the crime in the test story and to formulate an 'hypothesis' as to who committed the crime. The objectives set out for the users were:

- Using the software provided, try to extract from the text the key elements that make up the plot such as the main characters and outline of the story.
- Who committed the crime and how?
- Each character, fact and event extracted awards you points weighted by their importance. The aim is to score as many points as possible.

After the permitted hour of analysis, each user was allowed a maximum of 30 min to write a synopsis or hypothesis of the story, which could be in the form of a few paragraphs describing the whole story, or a paragraph for each character describing who they are and how they fit into the story. In order to compare and score the user's answers, a master marking sheet was drawn up containing all of the relevant facts (characters, relations and events) extracted from the story. Weights were then assigned to each item based on an assessment of its importance to the story and the difficulty in extracting it. This assignment of weights to information is necessarily a subjective process as different persons may assign different levels of importance to elements of the plot or have differing opinions as to the difficulty involved in extracting certain information. However, each of the user's answers are marked and scored against the same master marking sheet such that any skewness in the scores was uniform.

After completion of the test, all the units of information each user had extracted were added up to determine (a) the total number of units of information extracted and recorded by each user and (b) using the weights listed in the master marking sheet a weighted score intended to reflect the relative importance of the information extracted and recorded was calculated. Information extracted by a user but not recorded in their answer could obviously not be taken into account.

6. Results

The recall and score results of each group was characterised by its mean, standard deviation and number of data points. These values are shown in Table 4.

The recall of each group of participants shows the amount of information, or number of 'units of information' that were extracted as a percentage of the total number of units of information available for extraction. The score shows the

Table 4 – Results from the preliminary experiment.

	Model A	Model B	Search
Mean (recall)	0.24975	0.290169	0.275948
Std dev (recall)	0.0821294	0.108424	0.138394
Mean (score)	0.320643	0.329450	0.266382
Std dev (score)	0.117804	0.111308	0.100209
No. data points	4	4	4

relative importance of the information extracted by each group of participants expressed as a percentage. The recall and score totals for each group appear in Fig. 11.

No significance can be attached to the variation between the recall and score means of the three groups due to the large standard deviations. The wide standard deviations indicate that there were significant differences in performance for each person. This is to be expected in a test of this nature and is the result of varying abilities, effort, personality types, experience, and aptitude of the participants, inconsistent reporting and recording of findings by participants and familiarity with the type of subject material.

The wide standard deviations within each group coupled with the closely set means across groups indicate that there is a high degree of overlap in the information extracted by the three groups. This also prevents any significance being attached to the difference in the means of the three groups.

While it can be anticipated that there would be a greater dissimilarity between the results of each group as the time allowed for analysis is decreased, it is unlikely that the allowed analysis time of 1 h was too long. Had the analysis time of 1 h been too long one would have expected higher recall values for each group and lower standard deviations across the groups. In fact, no user achieved greater than 0.43 on the recall measure.

There are four categories of information that could be extracted from the novel: locations, characters, relations and personal facts, and events.

Locations are those entities that refer to the locations, towns or places. There are only three location references that are relevant to the story. Characters refer to entities of type person, of which there are 22 significant characters in the story. Types of relations to be extracted include familial relations (parents, children, and wives), professional occupations (doctors, attorneys), friendships and acquaintances. Personal facts cover a broad range of useful pieces of information about the characters that are relevant to the storyline, such as 'Poirot is a Belgian detective' and 'Dr Bauerstein is an expert in poisons'. There are 45 relations and personal facts about the characters that are relevant to the storyline. The final and largest category includes units of information describing what happened in the story, i.e. the storyline. There are 98 event units.

The mean score for each group in each of the four categories (expressed as a percentage) is shown in Fig. 12. The scores for each of the groups within each category were generally similar. The only category with scores greater than 50% was Characters.

6.1. *Locations*

The category with the largest variation in scores appears to be Locations. In this category the Search method appears to perform much better than the other two methods. It should be borne in mind, however, that there are only three location references. Of the participants using the Search method 50% found at most two locations and one participant did not find any. All of the relevant locations are also located together in the first few paragraphs of the text.

None of the participants using Model A found any locations. Investigation of Model A found that none of the locations occurred in the text often enough to meet the minimum frequency requirements imposed by Model A. The locations therefore would not have appeared in the graphical user interface with the result that the user would only have found location references if they happened to examine a sentence that is represented in the graph that contains a reference to the location. For example, the link for the nodes 'chemist' and 'strychnine' contains a sentence that states that the chemist's shop is in Styles St. Mary. However, because the location is not represented in the graph itself, the user is unlikely to place much value on the reference to the location. Therefore, the user is unlikely to extract references found in that manner.

Model B aims to compensate for the limitations of such frequency thresholds, by utilising NER. However, only 50% of those using Model B found at least one location. On examination it appeared that the poor accuracy of NE co-referencing had caused significant ambiguity. In this particular experiment it was found that the NE tagger could not properly differentiate and disambiguate the locations of the town Styles St. Mary and the residence Styles Court and they were tagged as a single entity. It was also found that one of the references to Styles St. Mary was allocated to the entity of Mary Cavendish. However, manual disambiguation of entities should be simple when the user considers the context of the

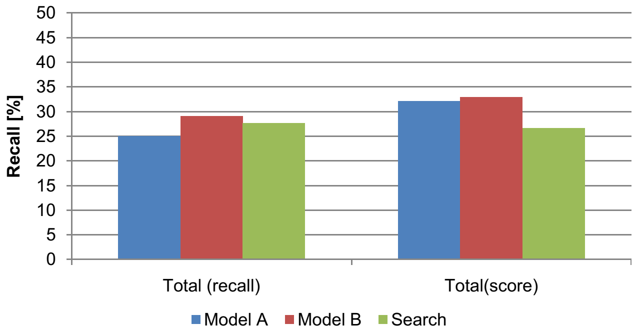


Fig. 11 – Mean results for each of the models.

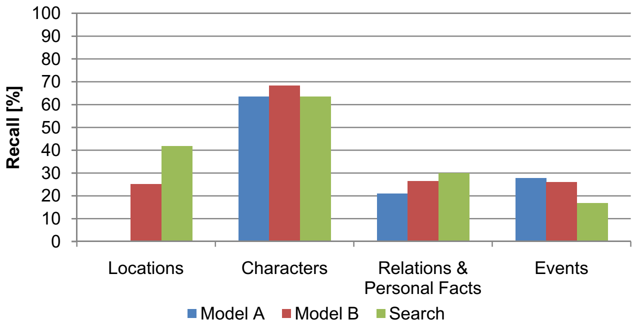


Fig. 12 – Results for each model across the categories.

references, by clicking on the node and reading the sentences displayed for the node.

The above indicates that when determining which extraction method would be most appropriate for the extraction of location information, one should carefully consider the nature of the text and the frequency with which location terms occur in the text. If the terms occur infrequently, they will fall below the threshold of Model A and will be excluded from the graph. Model B may then be the preferred method. However, if there are terms that are similar to other terms, some of them may be missed by Model B. While it is still difficult to find location terms using Search, in this case Search appears to have a better chance. For this type of information, a combination of methods may achieve the best results. However, as there were only three location terms to be extracted in this experiment, these results are inconclusive.

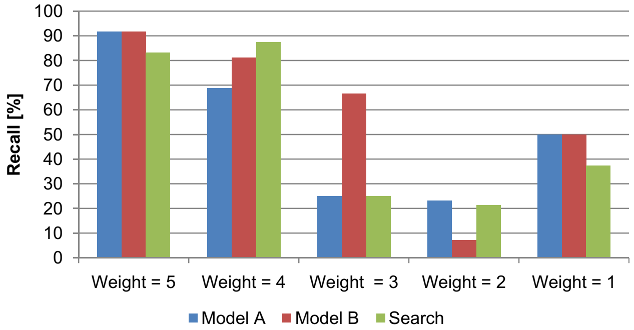


Fig. 13 – Percentage of characters extracted for each weight grouping.

6.2. *Characters*

All of the models performed significantly better in the Characters category than in any of the other categories. In this category the models performed very similarly, as illustrated in Fig. 12.

There are a number of reasons why characters were more easily found and extracted than other categories of information. In a novel, characters interact with each other which increases the likelihood of characters occurring near each other in the text. A user who finds one character is therefore likely to find another in addition. In a novel, the names of the characters are mentioned frequently and characters therefore score very well in a statistical based extraction method like Model A. The high frequency of characters and the likelihood of characters to appear in close proximity to each other in the text also assists the participants using the Search method, as the names are likely to appear in the text in close proximity to other terms that the user may be searching for. Once a character has been found, that character may also become a useful query term which assists the Search users to find other characters which interact with it. Characters, being named entities, are also assured of appearing in the Model B graphical user interface.

Since the objectives set for the users stated that they should extract the main characters from the story, the users probably put more effort into finding and extracting all of the characters and this information is probably also easier for a user to record.

It is therefore not surprising that users scored better in the characters category than in the other categories, regardless of the Model being used.

As described in the methodology, characters were assigned weights in the Master marking sheet based on their relative importance or contribution to the central plot of the novel. Characters can therefore be divided into weight groups according to the weight assigned to the character in the Master marking sheet. Fig. 13 illustrates the percentage of characters extracted for each weight group, where the most important characters have a weight of 5 and the least important characters have a weight of 1.

The figure shows that the recall for ‘key’ characters, or most important characters, is higher than the other weight groups. One reason for this may be that key characters, being important to the novel’s central plot, occur more frequently in the text than less important characters and therefore have a higher extraction rate. However, extraction of characters in the Weight 1 group does not fit this pattern. Examination of the characters in the Weight 1 group reveals that there are only two characters assigned to this group, namely Dorcas and Elizabeth Wells. Dorcas and Elizabeth are two of three housemaids in the story. They were assigned a weight of one, because they had little opportunity to be involved in the murder and thus, were not suspected. It is interesting to note that the character Dorcas had an exceptionally high extraction rate, an average of 75% across all of the participants, i.e. 75% of the participants extracted the character Dorcas. Since the assignment of weights is a subjective process, the high extraction rate is a strong indication that the weight chosen for Dorcas was too low. Assessment of the character Dorcas showed that she was frequently questioned by the detectives involved in the case, because she overheard some of the arguments between the victim and the suspects. Thus, re-evaluation of the weight assignment showed that the character Dorcas should have a weight of three. This would reduce the average extraction rate for the Weight 1 group to approximately 10%, which is inline with the expected extraction rate based on importance and frequency. Changing the weight assigned to Dorcas to three had a negligible effect on the extraction performance for the Weight 3 group.

Fig. 13 shows that Model A had a lower extraction rate than Search and Model B for characters in the Weight 4 group. It was found that the character Mr Hastings did not meet the threshold requirements of Model A to be included in the graph. Mr Hastings is the narrator in the book, and therefore references to him are most commonly and frequently referred to using “I”. Unfortunately, unless NER and co-reference disambiguation is performed, it is very difficult for users of Model A to manually disambiguate the character Mr Hasting with the narrator “I”, without referring to the full text. It was also found that the character Dr Wilkins, in the Weight 3 group, was excluded from the graph in Model A and therefore could not be extracted by any of the users of Model A. Dr Wilkins is referred to 19 times in the text and thus meets the frequency requirement for model A. However for inclusion in the graph, the node is required to be connected or related to another node and must meet the required support threshold for the link between the two nodes. Dr Wilkins did not meet the necessary support threshold for links to other nodes. It is interesting to note that only one of the Search users extracted the character Dr Wilkins, which can be attributed to the low extraction rate by Search users for the Weight 3 group.

Model B has a clear advantage over the other two models in extracting characters, due to its NER algorithm, which can be seen in its significantly better performance over both Model A and Search in the Weight 3 and Weight 1 character group. It is surprising to find that despite the automatic extraction and listing of NEs, Model B had an extraction rate of only 90% and 80% for the Weight 5 and Weight 4 groups respectively. Inspection of the results showed that three out of the four Model B users had extraction rates of 100% for both the Weight 5 and Weight 4 groups. The remaining user had extraction rates of 67% and 25% for these two groups respectively, which can be attributed to that user's evaluation of the importance of the characters to the storyline and/or the amount of effort he/she put into investigating and extracting characters and recording extracted results. In contrast the performance of Model A and Search for the Weight 5 and Weight 4 character groups was lower across multiple users and could not be attributed to just one user. The poor performance of Model B in the Weight 2 group was identified to be due to errors in the NER algorithm, whereby five of the seven characters in the Weight 2 group were not identified by the NER algorithm and thus were not included in the NE list.

All three models are effective at extracting characters, but it is expected that Models A and B would have a clear advantage over Search where the text being explored is not a novel and characters do not occur near each other in the text or interact. In this case, unless Search users know the names of the characters they are looking for, Models A and B would perform better, because of their automatic extraction methods. Model B would probably perform better than Model A because of its NER algorithm and the need in Model A to satisfy both a frequency threshold and a node interaction threshold. If one wanted to find characters only and nothing else (no relationships or events) NER could be used alone, and in that way might be more efficient. A person therefore needs to have an understanding of the nature of the text to be explored and the type of information required to be extracted in order to know which model is most appropriate to be employed.

6.3. *Relations and personal facts*

Relations by definition represent relationships and interactions and are therefore described by the manner in which they link characters; or characters and occupations, interests, nationalities, and facts about the characters. Relations cannot be extracted or retrieved using a search method in the same way as a character or a location can. Characters and locations, being nouns, can be extracted independently of the rest of the information in the text. A relation can only be extracted by describing the relationship between a minimum of two pieces of information, for example: "Dr Bauerstein is a friend of Mary's", "John practiced as a barrister". Therefore the extent to which one can extract relations is dependent on the extent to which one can extract relation independent information such as locations, characters, occupations and nationalities.

One possible explanation for relations having done poorly relative to characters is that users did not assign much importance to relations and did not expend much effort in extracting or recording them. That this might have been the case is supported by the fact that all three models did not perform well in this category.

This does not account, however, for the fact that search did better than either model in this category. Search probably outperformed the other models because the structure of a novel is such that when searching for characters one finds and reads the information about them, and thus one finds the relations. That this explanation is likely, is borne out by the fact that search does much better in relations than events, as is discussed below.

That Model A did not do well is perhaps expected as it does not focus on the links between NEs, it rather focuses on the links between frequent or important concepts.

If one wants to discover relations it may be most efficient to use a NER algorithm, followed by search, but this observation might only be true for novels and not for other forms of text.

6.4. *Events*

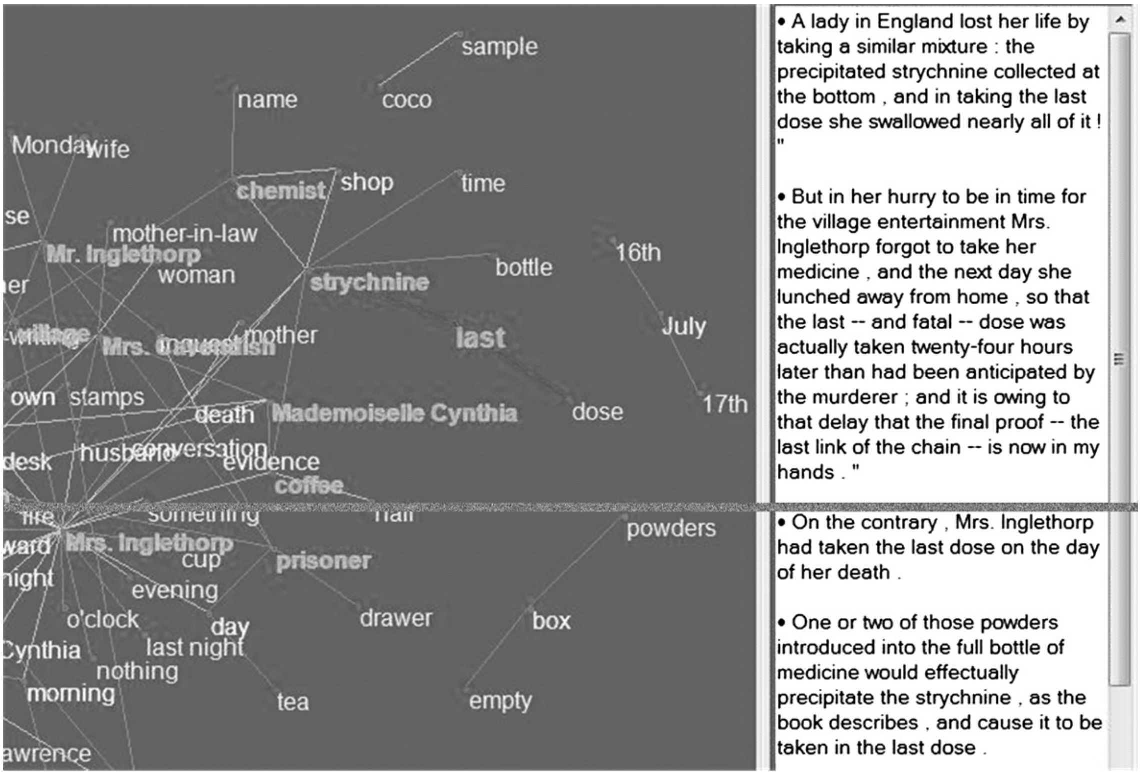
The events category is the largest category containing 98 'event-units' and is the only category in which the two evidence discovery models performed better than Search. Events are composed of one or more concepts and an associated action. This makes events harder to find and extract than characters or simple relations. The text contains numerous 'sideline' events that form sub-plots in the story, which make it difficult for users to find and identify the important events that solve the mystery. In the story the important events are widely distributed across the text and this makes it even more difficult to uncover these events using Search.

Additionally, the Search users had to think of and choose query terms to search through the text to find events, using new query terms extracted from their results to make further progress. This makes it difficult for a Search user to find the events that are relevant to the main storyline without reading the full text.

The evidence discovery users had the advantage of the graphical interface which in effect provided them with 'prepopulated search terms' in the form of the nodes in the graphs. Events may also be found more easily by the evidence discovery models because the concepts or characters in the events are connected by common nodes.

Model A connects important concepts and characters in a graph. Because events tend to revolve around important concepts and characters, these nodes have several connections and can be easily found in the graph. Fig. 14 shows an analysis of the links 'strychnine - last - dose' in the graph produced by Model A. The example in Fig. 14 shows how the following two very important pieces of information can be found:

- "One or two of those powders introduced into the full bottle of medicine would effectually precipitate the strychnine, as the book describes, and cause it to be taken in the last dose."



- A lady in England lost her life by taking a similar mixture : the precipitated strychnine collected at the bottom , and in taking the last dose she swallowed nearly all of it ! "

- But in her hurry to be in time for the village entertainment Mrs. Inglethorp forgot to take her medicine , and the next day she lunched away from home , so that the last -- and fatal -- dose was actually taken twenty-four hours later than had been anticipated by the murderer ; and it is owing to that delay that the final proof -- the last link of the chain -- is now in my hands . "

- On the contrary , Mrs. Inglethorp had taken the last dose on the day of her death .

- One or two of those powders introduced into the full bottle of medicine would effectually precipitate the strychnine , as the book describes , and cause it to be taken in the last dose .

Fig. 14 – Analysis of events in Model A.

- "But in her hurry to be in time for the village entertainment Mrs. Inglethorp forgot to take her medicine, and the next day she lunched away from home, so that the last - and fatal - dose was actually taken 24 h later than had been anticipated by the murderer; and it is owing to that delay that the final proof - the last link of the chain - is now in my hands."

Model B is designed around NEs, and focuses on the events connected to the NEs. Similarly to Model A, events can be easily found by exploring the graphs. Fig. 15 shows an analysis of the node 'Inglethorp' in the graph produced by Model B. The example in Fig. 15 reveals an important event between 'Mrs Cavendish' and 'Mrs Inglethorp':

- "Mrs Cavendish administered a safe, but effectual, narcotic to both Mrs Inglethorp and Mademoiselle Cynthia."

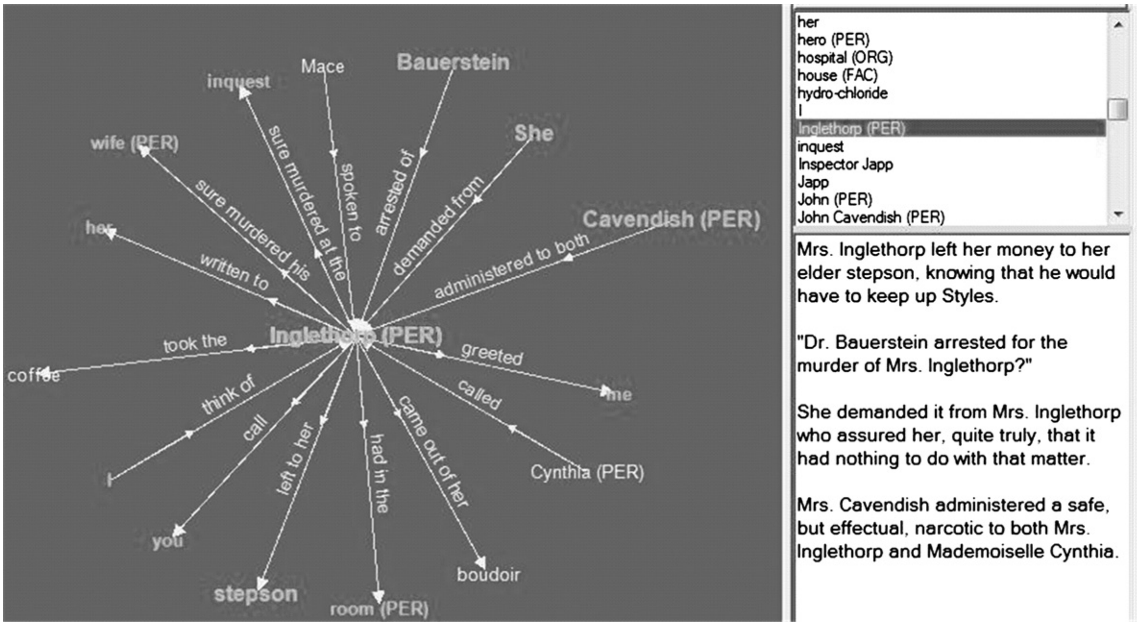
The evidence discovery models are best suited to the extraction of events and perform better than Search. In a small body of text one would expect a reading of the text or even perhaps searching to be adequate, equal to or perhaps even better than Models A or B, but this test shows that the longer and more information rich the text, the better the evidence extraction models will perform in comparison to Search.

6.5. *Who committed the crime and how*

It is interesting to analyse how effective the three models were in solving the murder mystery in the text. It is important to remember that the final chapter of the novel was excluded from the text used in the experiment. This ensures that the conclusion which reveals the details of who committed the crime, how it was done and the murderer's motive was not given to the users. However, the nature of an Agatha Christie novel ensures that the reader is given all of the clues in the preceding chapters. In order to identify the criminal(s) and put together the pieces of how the crime was executed, the user is required to make deductions and decide which clues are fact and which are misleading 'red-herrings'.

In order to evaluate the second objective set for the users ('who committed the crime and how?'), each fact which contributes to the answer was put into one of three sets, namely Who, How, and Why. This required the user to have successfully extracted information from each of the four categories. Each of the Who, Why and How sets can accordingly be assigned a total absolute score calculated as the sum of the individual scores of the elements of that set, whether a location, character, relation or event. Fig. 16 illustrates the mean score expressed as a percentage of the maximum possible score for each of the Who, Why and How sets for each method. The "Total" indicates each method's mean score across all three plot elements. It is evident from Fig. 16 that in each set Model A performed best, followed by Model B and finally search. This mirrors the results for the Events category discussed above.

That the performance of each of the methods in extracting the plot mirrors their performance in extracting events is to be expected. Although a plot is comprised of locations, characters and relations as well as events, unless a user understands



her
hero (PER)
hospital (ORG)
house (FAC)
hydro-chloride
I
Inglethorp (PER)
inquest
Inspector Japp
Japp
John (PER)
John Cavendish (PER)

Mrs. Inglethorp left her money to her elder stepson, knowing that he would have to keep up Styles.

"Dr. Bauerstein arrested for the murder of Mrs. Inglethorp?"

She demanded it from Mrs. Inglethorp who assured her, quite truly, that it had nothing to do with that matter.

Mrs. Cavendish administered a safe, but effectual, narcotic to both Mrs. Inglethorp and Mademoiselle Cynthia.

Fig. 15 – Analysis of events in Model B.

the events, they will have a list of disparate and discrete locations, characters and relations, and no sense of how these contribute to the whole. It is only through extracting events that a user can piece together a plot. It is therefore to be expected that the method that performs best at extracting events would also perform best at enabling a user to extract a plot.

6.5.1. *Who*

There were two people involved in the murder of Mrs Inglethorp. While five out of the twelve participants (approximately 40%, two Model A users, one Model B user, and two Search users) identified the main culprit (Miss Evelyn Howard), who planned and executed the murder, only one user was able to identify the accomplice, Mr Alfred Inglethorp. In order for the users to discover Miss Howard's involvement in the murder, the users needed to find the portion of the text that reveals the contents of a letter that Mrs Inglethorp found. The characters in the book had different ideas as to the contents of the letter, but it was found to be an unfinished letter of correspondence between the two culprits, who were planning the murder. To deduce who the accomplice was, the user needed to discover that Miss Howard and Alfred Inglethorp were 'distant relations' and perhaps conclude that despite the pretence of them hating each other, they in fact were planning on eloping together. The fact that the text contained several interpretations of the letter and that all three methods performed poorly at relation extraction accordingly made discovery of both murderers and an understanding of the relationship between them difficult.

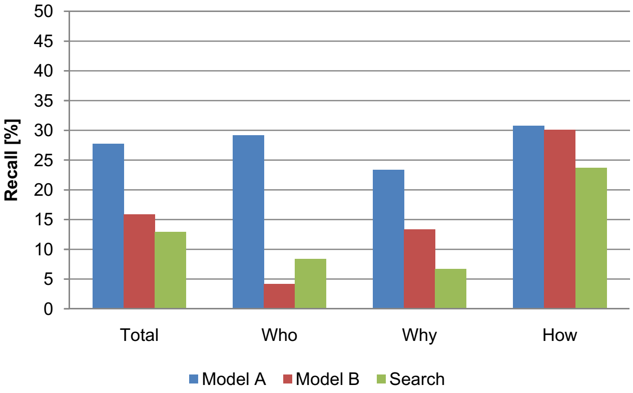


Fig. 16 – Analysis of results for Objective 2: 'who committed the crime and how?'.

6.5.2. *Why*

Approximately 50% of users for each model discovered that, on Mrs Inglethorp's death, Alfred Inglethorp would inherit her money. However, only two of the users, both using Model B, stated that the money from Mrs Inglethorp's will was the motive for the murder, despite the fact that neither of these users stated that Alfred Inglethorp was the murderer. Alfred's title to inherit the money coupled by Evelyn Howard being identified by most users as the murderer created confusion for the users, as these characters tried to hide their relationship by pretending to hate each other. Examination of the motive for the murder would have enabled the users to connect the inheritance of Mrs Inglethorp's money and the discovered fact that the murderer had an accomplice. In order for Miss Howard to benefit from the murder she had to have an agreement or arrangement with Alfred Inglethorp. Only one user, using Model A, deduced that a probable romantic relationship existed between Alfred Inglethorp and Miss Howard. This again demonstrates the importance of a method that performs well at extracting relations.

6.5.3. *How*

From Fig. 16, it can be seen that Model B and Search perform considerably better in this set than in the other two sets. This is probably because the ‘how’ does not require an understanding of the relations and how all the pieces of the story fit together. The ‘how’ only considers events, which are directly concerned with how the murder was accomplished. Thus the performance of the models is determined by their ability to extract events from the text.

7. Discussion

Most common problems usually associated with forensic computing are the complexity and quantity of data. Traditional data analysis methods rely on search methods to find potential evidence, which requires the analyst to know what he/she is looking for. This research explores a different and novel approach of analysing linguistic textual data using unsupervised methods. However, when investigating a real dataset one would not want to restrict the analysis to only a portion of the data. A typical dataset from a digital investigation would include numerous entities such as documents, operating system files, deleted entities, page files and more. Textual artefacts collectively referred to as documents are very important to many digital investigations (Beebe et al., 2007; McCue, 2006) and include emails, reports, letters, notes, meeting minutes, text messages etc. The different types that are encountered in investigations require different methods of analysis to find the most relevant information.

For example, if one wished to examine an email dataset such as the Enron email dataset[3] one would require analysis methods to cater for both the structured textual data in the email headers as well as the unstructured linguistic text in the body of the emails. An additional pre-processing step is therefore needed to separate the data into its different types for pre-processing and analysis, and a post-processing step is required to merge the data back into an undivided and complete summary of the data. The information retrieved from the headers or any file meta-data and system data could be incorporated into the textual graphs as a type of meta-overlay indicating another level of named entities, relations and timestamps. By extending the evidence discovery framework to use different data processing models to cater for different types of files, documents and data, the evidence discovery system may be applied to real datasets that typically include computer hard drives containing large amounts of unknown entities. Incorporating additional meta-data may enable the extracted data to be classified and clustered and presented to the investigator in a hierarchical manner which would enable the investigator to see an overview of the data and then drill down into the details of the data, which would be necessary for the analysis of large datasets.

An analysis using the evidence discovery framework presented in this paper performed on the Enron email dataset could provide valuable insight into the topics of discussion among groups of employees, creating links based on the content of the emails and not just based on the relationships created from the sender and receiver information. Such an analysis and the adaptations to the framework discussed above present worthwhile areas of research.

8. Conclusion

The aim of this research was to investigate the adaptation and application of text mining methods to the analysis of textual data for the purposes of evidence discovery. It was hypothesized that information extraction techniques combined with visual exploration techniques can assist in identifying suspects and events, and the relations between these entities which could assist an investigator to: piece together the story surrounding a crime, create hypothesises or potential leads for further investigation, or identify pieces of data which could form useful supporting evidence in a trial.

A novel framework, in the form of an evidence discovery system, was proposed in which to perform evidence discovery in an attempt to meet the challenges identified. By utilising unrestricted, unsupervised information extraction techniques, the investigator does not require input queries or keywords for searching, thus enabling the investigator to analyse portions of the data that may not have been identified by keyword searches.

The evidence discovery system produces text graphs of the most important concepts and associations extracted from the full text to establish ties between the concepts and provide an overview and general representation of the text. Through an interactive visual interface the investigator can explore the data to identify suspects, events and the relations between suspects. This assists the investigator to piece together the story surrounding the crime, create hypothesises for potential leads for further investigation, or identify pieces of data which could form useful supporting evidence in a trial.

Two models were presented for performing the relation extraction process of the evidence discovery framework. Model A takes a statistical approach to discovering relations based on co-occurrences of complex concepts. Model B utilises a linguistic approach using NE extraction and IE patterns. By working within a framework, individual components can be worked on and improved in a comparable manner.

A preliminary study was performed to assess the usefulness of a text mining research approach to evidence discovery as against the traditional IR approach. The results produced by each of the relation extraction models used within the evidence discovery framework were shown to be useful to the users in the experiment. The novel approach to text analysis for evidence discovery presented in this paper is therefore a viable and promising approach. The preliminary experiment showed that the results obtained from the evidence discovery system, using either of the relation extraction models, are sensible and useful. It is therefore concluded that this framework may be helpful in digital forensics.

[3] The Enron email dataset contains data from about 150 users, mostly senior management of Enron, organised into folders. The corpus contains a total of about 0.5 M messages. This data were originally made public, and posted to the web, by the Federal Energy Regulatory Commission during its investigation. The dataset is now available from the Carnegie Mellon University School of Computer Science at http://www.cs.cmu.edu/~enron/.

9. Future work

The relative newness of the field of digital forensics means that the scope for future research is wide open. There are many fields that are producing interesting research, with much promise for cross-disciplinary application, such as link analysis, social network analysis, temporal text mining, and temporal text analysis.

In order to gain a clear insight into the usefulness of the evidence discovery framework, a case study should be performed on a real dataset, such as the Enron email dataset discussed above. This would identify any problems of practical implementation of the framework and provide the best insight as to how the framework and its components should be worked on and improved. The framework could also be adapted as discussed above to incorporate additional metadata to the extracted data, which can possibly be used to classify and cluster the data to be presented to the analyst in a hierarchical manner which would enable the analyst to see an overview of the data and then drill down into the details of the data. A case study of this nature would clarify the needs and requirements of the users to improve their analysis of the data, in order to make the investigation process more accurate and efficient.

An investigation of theme extraction could produce a promising extension to the evidence discovery framework. Once relations have been discovered, extracted, and linked to form graphs, graph analysis techniques could be applied to recognize themes among the relations. Themes may possibly present themselves in the graphs as long threads of connected nodes, or circular or other patterns of connected nodes. Extracted themes presented to the user could assist the user to focus on, or rule out certain portions of the text to be analysed.

The Internet and search engines have brought search technology to all computer users. Most users are familiar and well practiced with search technology, which makes them more efficient utilising search tools as opposed to the newer and unfamiliar discovery tools. It is thought that a best of ‘both worlds approach’ could easily be achieved by combining search-based analysis methods with discovery methods. Search functionality could be easily incorporated in the discovery system to enable the user to search for nodes within the text graphs and to search for items in the source text that were found in the graph.

Acknowledgements

This work was supported in part by grants from the Council for Scientific and Industrial Research (CSIR), South Africa and the National Research Foundation (NRF), South Africa.

REFERENCES

Abraham, T., 2006. Event sequence mining to develop profiles for computer forensic investigation purposes. In: ACSW Frontiers ’06: Proceedings of the 2006 Australasian workshops on Grid computing and e-research. Australian Computer Society, Inc., Darlinghurst, Australia, Australia, pp. 145–153.

Beebe, Lang Nicole, Clark JG. Digital forensic text string searching: Improving information retrieval effectiveness by thematically clustering search results. In: Digital investigation, vol. 4. Elsevier Ltd.; 2007. 49–54.

Brill E. A simple rule-based part of speech tagger. In: Proceedings of the Workshop on Speech and Natural Language. Morristown, NJ, USA: Association for Computational Linguistics; 1992. pp. 112–116.

Charniak, E., 08 2005. The charniak parser (nlparser). ftp://ftp.cs.brown.edu/pub/nlparser/.

Chen H, Chung W, Xu JJ, Wang G, Qin Y, Chau M. Crime data mining: a general framework and some examples. Computer 2004;37(4):50–6.

Churcher, N., Irwin, W., Cook, C., 2004. Inhomogeneous force-directed layout algorithms in the visualisation pipeline: from layouts to visualisations. In: Proceedings of the 2004 Australasian Symposium on Information Visualisation. Australian Computer Society, Inc., Darlinghurst, Australia, Australia, pp. 43–51.

de Waal, A., Venter, J., Barnard, E., 2008. Applying topic modelling on forensic data: A case study. In: Proceedings of the 4th Annual IFIPWG 11.9 International Conference on Digital Forensics. Kyoto, Japan, Springer.

Dozier C, Jackson P. Mining text for expert witnesses. IEEE Software 2005;22(3):94–100.

Eades PA. A heuristic for graph drawing. In: Congressus Numerantium, vol. 42; 1984. 149–160.

Fan W, Wallace L, Rich S, Zhang Z. Tapping the power of text mining. Commun ACM 2006;49(9):76–82.

Fellbaum C, editor. WordNet. An electronic lexical database. MIT Press; 1998.

Greenwood, M. A., Stevenson, M., Guo, Y., Harkema, H., Roberts, A., 2005. Automatically acquiring a linguistically motivated genic interaction extraction system. In: Proceedings of the 4th Learning Language in Logic Workshop.

Grishman, R., 2007. Jet (java extraction toolkit). (Last accessed 10. 04.07). URL http://cs.nyu.edu

Guernsey L. Digging for nuggets of wisdom. The New York Times, http://query.nytimes.com/gst/fullpage.html?res=950CE5DD173EF935A25753C1A9659C8B63&sec=&;spon=&pagewanted=1; October 2003 (Last accessed 17.06.08).

HeliosSoftwareSolutions. Textpad 4.4.0, www.TextPad.com; 2000.

Hotho A, Nrnberger A, Paa G. A brief survey of text mining. LDV Forum - GLDV. J Computational Linguistics Lang Technol MAY 2005;20(1):19–62.

Infogistics, 2001. Nlprocessor – text analysis toolkit. (Last accessed 19.11.08). URL http://www.infogistics.com/textanalysis.html.

Jurafsky D, Martin JH. Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition. 2nd ed. Pearson Prentice Hall; 2008.

LDC. ACE (Automatic content extraction) English annotation Guidelines for entities. Linguistic data Consortium, University of Pennsylvania, http://www.ldc.upenn.edu/Projects/ACE/; 2005. v. 5.6.1 2005.05.23.

Lebert, M., 2008. Project Gutenberg (1971-2008). Project Gutenberg, eText-no. 27045. URL http://www.gutenberg.org/etext/27045.

Louis, A., 2009. Unsupervised discovery of relations for analysis of textual data in digital forensics. Master’s thesis, University of Pretoria.

Manning CD, Schütze H. Foundations of statistical natural language processing. Cambridge, MA: MIT Press; 2001.

McCue C. Data mining and predictive analysis: Intelligence gathering and crime analysis. Elsevier; 2006. 368 pp.

Meyers, A., Grishman, R., Kosaka, M., Zhao, S., 2001. Covering treebanks with glarf. In: Proceedings of the ACL 2001

Workshop on Human Language Technology and Knowledge Management. Association for Computational Linguistics, Morristown, NJ, USA, pp. 51–58.

Patton MQ. Qualitative research and evaluation methods. 3rd ed. SAGE; 2002.

Salton G, McGill MJ. Introduction to modern information retrieval. McGraw-Hill; 1983.

Sekine, S., 2006. On-demand information extraction. In: Proceedings of the COLING/ACL on Main conference poster sessions. Association for Computational Linguistics, Morristown, NJ, USA, pp. 731–738.

Surdeanu, M., Harabagiu, S. M., 2002. Infrastructure for open-domain information extraction. In: Proceedings of the 2nd International Conference on Human Language Technology Research. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp. 325–330.

Swanson DR. Two medical literatures that are logically but not bibliographically connected. J Am Soc for Inf Sci 1987;38(4): 228–33.

Swanson, D. R., 1991. Complementary structures in disjoint science literatures. In: Proceedings of the 14th Annual International ACM/SIGIR Conference. pp. 280–289.

Please cite this article in press as: Louis AL, Engelbrecht AP, Unsupervised discovery of relations for analysis of textual data, Digital Investigation (2010), doi:10.1016/j.diin.2010.08.004