
http://dx.doi.org/10.1109/SMC.2016.7844519

Final accepted version (with author's formatting)

Available from Middlesex University’s Research Repository at http://eprints.mdx.ac.uk/21191/

Copyright:

Middlesex University Research Repository makes the University's research available electronically.

Copyright and moral rights to this thesis/research project are retained by the author and/or other copyright owners. The work is supplied on the understanding that any use for commercial gain is strictly forbidden. A copy may be downloaded for personal, non-commercial, research or study without prior permission and without charge. Any use of the thesis/research project for private study or research must be properly acknowledged with reference to the work's full bibliographic details.

This thesis/research project may not be reproduced in any format or medium, or extensive quotations taken from it, or its content changed in any way, without first obtaining permission in writing from the copyright holder(s).

If you believe that any material held in the repository infringes copyright law, please contact the Repository Team at Middlesex University via the following email address:

eprints@mdx.ac.uk

The item will be removed from the repository while any claim is being investigated.
Associative Search through Formal Concept Analysis in Criminal Intelligence Analysis

Nadeem Qazi∗, B.L. William Wong∗, Neesha Kodagoda∗ and Rick Adderley**
∗Interaction Design Centre, Middlesex University,
London, United Kingdom
∗∗AES Solution Ltd, UK
n.qazi@mdx.ac.uk, w.wong@mdx.ac.uk, N.Kodagoda@mdx.ac.uk, RickAdderley@a-esolutions.com

Abstract—Criminal Intelligence Analysis often requires a search different from the semantic and keyword based searching to reveal the associations among semantically and operationally connected objects within a crime knowledge base. In this paper we introduce associative search as a search along the networks of association between objects like people, places, other organizations, products, events, services, and so on. We also propose an associative search model based on the 5WH associated concepts of a crime, i.e. WHAT (what has happened), WHO (who was involved in the crime), WHEN (the temporal information of the crime), WHERE (the geo-spatial information of the crime) HOW (the modus-operandi used in committing a crime). We have employed Formal Concept Analysis theory to reveal the associations, highlighting Hot Spots, offender’s profile and its associated offenders in a criminal activity.

Keywords—criminal intelligence; association rules, semantic search; formal concept theory; linked analysis

I. INTRODUCTION

In our investigations of how analysts think and reason about problems in criminal intelligence analysis, we have observed a practice we refer to as “associative questioning”[1]. This is the practice where an analyst asks a variety of questions to learn more about the diverse nature of the context in which the crimes were committed. This context is important for making sense of the situation that would help solve the crime. For example, some analysts apply the 5WH (Who, What, When, Where, Why and How) analytic reasoning model to ask questions that help them, for example, to discover who else might have been involved in the crime, what other factors or events could be relevant, when these and other events occurred, why or what have or could have motivated the act. They may also seek to understand the motivations of the crime by reasoning with the data through the perspective of crime theories such as the Crime Triangle [2], and also how was the crime and other similar crimes committed? Were there similarities between past and current crimes, e.g. is this part of a trend? e.g. could this be the same person re-offending? Currently, to collate such a comprehensive picture of the crime requires the analyst to painstakingly perform directed searches of many different databases. This takes up a lot of time and effort.

The purpose of this paper is to report on our efforts to develop an alternative search method to support the “associative questioning” process – associative search. It is intended to unfold the operational association across the network of data objects to retrieve operationally significant associations, and not just data that are semantically related. The next section provides some background to the problem domain of Criminal Intelligence Analysis and similar work on Associative Searching. We also report on our attempts to introduce Formal Concept Analysis (FCA) as a possibility for reducing the rigidity of the 5WH model as the basis for making the associations.

II. BACKGROUND

Intelligence analysis in general, is the process of collecting, reviewing and interpreting a range of data [3], and intellectually distinguishing the “significant from the insignificant, assessing them severally and jointly, and arriving at a conclusion by the exercise of judgment: part induction, part deduction”[4], and “.. part abduction” [5].

Criminal intelligence analysis is defined as the “… study [of] criminals, crime suspects, incidents, issues and trends ... [to] identify relationships or connections between different crimes in different places” [6]. Criminal intelligence analysis can generally be divided into tactical, operational and strategic analysis. Tactical analysis focuses on supporting day to day street level investigations, whereas operational analysis generally focuses on supporting investigations, identifying links between suspects, and their involvement in crimes and criminal activity, and developing profiles of known or suspected criminals. Strategic analysis is generally focused on the study of long term trends and patterns to inform higher level decision makers of threats and emerging crime issues. This is used to advise on allocating resources to different types of crime, or increasing training in a particular crime-fighting technique [6]. In some police forces operational and strategic analysis overlap, while in others, these functions are performed in separate departments.

Analysts face a number of significant difficulties in the information analysis process, including making sense of data from multiple sources that need to be collated and organized into meaningful ways that can lead to sense-making and insight generation. Data is also of varying quality and reliability, out of sequence, lacking context, and missing, ambiguous and uncertain. Analysts may also be working on several cases at the same time and may be difficult to distinguish the relevance or similarities among these cases.

Associative search may be defined as search along the networks of associations between objects such as people, places, other organizations, products, events, services, and
so forth. It is different both from keyword and semantic based search techniques. For example, keyword based searching does not consider the meaning of the given query and matches exact words from related documents. Semantic based search, on other hand looks for the content that matches the meaning of the question, however it still does not leverage the power of associations of concepts in the search domain. However a similar data analysis technique known as Link Analysis in the literature, which in contrast, uses graph theory to build network of interconnected objects in order to explore patterns and trends. It evaluates relationships among various types of nodes (objects), including organizations, people and transactions (connections) between nodes and visualizes it through time and event charts, association and activity matrices, and Link Analysis Diagrams.

Several associative search definitions are reported in the information retrieval literature. For example Raaijmakers et al [7] has defined associative search as the cue-dependent retrieval system of human interconnected memory. Spivack [8] explains in his blog that associative search does not merely see the import of the question; it should also interpret and can reason about relationships in the data. Oh and Cho [9] has defined the associative search as a human retrieving memory process that can be implemented through constructing a semantic network consisting of related objects. They have demonstrated it through an interactive visualization of a semantic network of mobile log, consisting of data, such as GPS, Call, SMS, picture viewer, MP3, charging and photo tagging. In addition to this, statistical, contextual semantic measures and ontology have also been used in literature for determining associations in textual information [10, 11]. Moawad et al [12] have demonstrated an ontology based Arabic semantic search engine, consisting of a syntactical search engine, an interactive semantic query analyser and semantic ranker. The interactive semantic query analyser, however, incorporates the associative search, by suggesting associated concepts to the given input query through a domain ontology, and then through a syntactic search engine (Google API), finds and ranks the document that matches these associated concepts. In a software engineering domain, Takuya and Masuhara [13] have developed a source code recommendation tool based on an associative search engine called GEETA. It spontaneously searches and displays example programs while the developer is editing a program text.

The critical issue in associative search is to establish the scope or levels of associations among the related concepts over the given data set. This depends on the understanding of user intention behind the query. Query expansion, relevance feedback and pseudo-feedback are recommended methodologies quoted in the literature to understand the user intention behind a query. A new algorithm for query expansion based on the distance constraint activation model of human memory has recently been proposed [14].

In addition to understanding the user intention behind the query another important thing is the establishment of the associations between related concepts. Google has used the idea of the Knowledge Graph to represent the association between all the real world objects. The purpose of Knowledge Graph is to find the right information, get the best summary available and go deeper and broader into the content all in order to return more relevant information to the user and to better understand the content the user is looking for. In order to understand the underlying goal of the user query, Google also uses the browsing history of the user and also consider what other people have been searching on the same or similar issues. The social web site LinkedIn while performing associated search uses a patented relevance algorithm to calculate the relevance score, which in turn is used to establish the association among the people. Amazon determines the association between the people with the help of user profile and a recommendation system. Facebook uses semantic search through a Graph Search algorithm to find information from within a user's network of friends.

In this paper, we have used a different approach to answer these issues, and have proposed to use Formal Concept Analysis (FCA) lattice to visualize the query scope of the given knowledge base. We also have demonstrated the use of frequent item set mining and A Priori algorithm to determine the association rules between the linked concepts in the data.

III. ASSOCIATIVE SEARCH MODEL IN VALCRI DOMAIN THROUGH FCA

In proposing our method we have focused towards the criminal data related to burglary offence. A crime such as a burglary offence usually contain temporal and spatial information, i.e. when and where a burglary offence has occurred, and with what modus of operandi used by the offender to commit the offence. The analytical reasoning process in analysing and solving these types of offences needs relevant information for constructing arguments and making judgments. An analyst might be interested to know about what was found at the place, left by the offender, or he might also be interested in knowing the potential list of offenders for a particular burglary offence, helping him achieve some underlying goal. These types of queries go beyond the functionality of semantic and keyword based search as they need to establish the associations between the connected objects in a knowledge base (KB).

Answering these types of queries and inspired by the Crime Triangle and the Routine Activity Theory [2], we have proposed an associative search model, implemented using Formal Concept Analysis consisting of five general though associated concepts: WHAT (what has happened i.e. type of offence), WHO (who has committed the crime, identifying possible suspect(s) e.g. by criminals who might share a similar modus of operandi or share a similar spatial temporal area of operation), WHEN (when the offence took place), WHERE (the location about the offence), WHY (what was the cause for this offence to happen), and HOW (the modus-operandi used in the committing a crime). Each of these 5WH concepts represent a question and are also linked with each other through a further set of properties or attributes. The objective is to find the connected information between WHO, WHAT, HOW and WHY temporally and spatially (WHEN & WHERE), and generating association rules among all these concepts. For example, it should answer the general questions such as:

- Who are the known offenders operating in an area based on how they commit crime (Modus Operandi)?
- What is the Modus Operandi used by an offender to commit a crime?
• Show crime trends and patterns through spatial-temporal and modus operandi information.
• Summarise the number of times an offender has committed a crime, in his areas of operation.

The discovery of these conceptual clusters in the database is non-trivial, however lattice theory through the framework of Formal Concept Analysis (FCA) can bring more formal mathematical thinking to assist in discovering and representing these type of concepts in terms of the context of the data, and hence we have focused on FCA to implement our proposed search model. In the next section, we give a brief description of FCA followed by the methodology.

IV. FORMAL CONCEPT AND ASSOCIATION RULES

Formal Concept Analysis [15], is a data analysis technique grounded in Lattice Theory that has been extensively applied for the purpose of knowledge discovery, in the fields of psychology, sociology, anthropology, medicine, biology, linguistics, computer sciences, mathematics, industrial engineering web mining and recently in crime analysis [16]. It represents the subject domain through a formal context made of objects and attributes of the subject domain. Philosophically a concept is a unit of thoughts consisting of extensions and intensions. Wagner [17] has coined the term extension to all the objects belonging to a concept and intension to all attributes valid for all those objects. In other words a concept is constituted by its extension, comprising all objects belonging to this concept, and its intension, comprising all attributes (properties, meanings) that apply to all objects of the extension. The set of objects and attributes, together with their relation to each other form a formal context, which can be represented by a cross table called as formal concept table. Each row in the Formal concept table corresponds to objects, each column corresponds to attributes, and a binary value denotes the relationship between them. One of the major outputs of this cross table is a concept lattice reflecting generalization and specialization between the derived formal concepts. In the lattice structure, formal concepts are represented by the nodes which have attributes placed over the nodes and objects under the nodes. To retrieve extensions, one must simply trace all paths leading down from the node of the attribute of interest. Similarly, it is possible to retrieve intensions by tracing all paths leading up from the node representing the object of interest.

V. REPRESENTING CRIMES THROUGH FCA

We have proposed to use the notion FCA consisting of the formal concept as the basis unit for visualizing the SWH questions of our associative search model with each concept of the FCA lattice representing a question through its intent (attribute) and the answer of the question through its extent (object) or vice versa. The adjective “formal” is used to emphasize that these are formal notions that may not necessarily consist of real objects and rather contains object-like items as formal objects having their features or characteristics as formal attributes. A set of related questions (either specific or general) thus can be visualized in an organized hierarchical scheme of formal concept under a lattice structure. This hierarchy may be useful as a querying interface to browse a set of possible questions, together with their answers from a dataset at varying levels of granularity and specificity to the user. Each of these sets represents the answers to a particular question, which is characterized by the properties shared by the elements of the set. In this study, we have used both general and specific questions specifically focusing towards those questions for which the objects are the answers in a given context, for example:

• Who else has committed similar burglary offences
• What are other/similar criminal networks of an offender for similar offences/operation/spatial and temporal) and Modus of operandi?
• Where else has he/she been spotted in a crime, highlighting Crime Hot Spot (CHS) of an offender or crimes?
• Were there any similarities between past and current crimes associating its temporal, spatial and other similar criminal activities, thus needing to query the Geo-Spatial profile of the offender? It however, stems from more specific questions such as:
  o How offenders are organized by area and modus of operandi.
  o What happened on given day e.g.(Sunday) in an area e.g. High Street?
  o Where an offender mostly likely to commit a type of crime using a specific modus of operandi? Who else had committed the same offence in that location?

However, FCA can generate thousands of questions, making browsing the hierarchy cumbersome, therefore in implementing associative search we have visualised these queries in three separate tempo-spatial formal contexts as illustrated in Table I, generating three tempo-spatial lattices. Formal context as mentioned before is a subset of the cross product between its sets of objects and attribute. Mathematically, it is defined as a triple K := (G,M,I) with M a set of attributes of a question and G represent sets of objects as answers and I ⊆ G×M is binary relation defined between G and M.

A. Offender Profile

To represent WHO type questions of our associative search model we encapsulated set of offenders (G) as

<table>
<thead>
<tr>
<th>Geo Spatial and Temporal Formal Concepts</th>
<th>Purpose</th>
</tr>
</thead>
<tbody>
<tr>
<td>Objects</td>
<td>Attributes</td>
</tr>
<tr>
<td>1 Offenders</td>
<td>Street, Offences, Month, Day and Time of Offence occurred.</td>
</tr>
<tr>
<td>2 Offenders</td>
<td>Crime, Ref, Offences, Month, Day and Time of Offence</td>
</tr>
<tr>
<td>3 Geo Spatial Location (Streets) of offence</td>
<td>Offender, Offences, Month, Day and Time of Offence occurred</td>
</tr>
</tbody>
</table>

objects or extents and their spatial, temporal information as intents (M) of the formal context. This formal context thus represents the geographical and temporal profile of the offenders. The formal context was made using the data consisting of two offence types, i.e. Burglary Other Building and Attempt Theft of Motor Vehicle along with spatial and temporal information such as street, month day and time of the offence occurred. The time of the offence
is a many-valued attribute having any value within 14 hours of the day. We therefore used the idea of conceptual scaling to transform this attribute into its symbolic value which resulted into four periods of the day: morning (from 6 a.m. to 12 a.m.), afternoon (from 12 a.m. to 6 p.m.), evening (from 6 p.m. to 12 p.m.), and night (from 12 p.m. to 6 a.m.). All the formal concepts of this formal context were then clustered in a binary relation in a cross table (Table II) representing X (true) in cell (i, j) of object i with attribute j, and an empty cell (false) otherwise. The FCA lattice was then generated from this cross table using Lattice Miner and is shown in Figure 2, representing formal concepts through circles, with all their intents and extends mentioned by blue and red text respectively. The top most concepts (FC1) of the lattice (Figure 2) contain all objects in its extension, whereas the bottom concept (FC15) contains all attributes in its intensity. The color intensity of each node reflects the number of the objects counts in the node. All the objects in each of the formal concepts share all the attributes in that formal concept. This means in Figure 2 the concept (FC5) bear offenders Clouser, Horethm and Khong, all of them were found to commit the offence ‘Burglary Other Building’ in February on Sunday night at Stonnal Road.

The generated lattice of the Figure 1 shows a hierarchy of the both general and specific questions with the intent of each concept and their answers in the extents of the same concepts. In general every concept of an FCA concept will inherit all attributes associated with its super concepts. This means the object of the concept will bear all the connecting attributes appear on the upward leading path from the concept. Similarly an attribute owns all the connecting objects on a downward path from its concepts. For example, starting from the general question the top most concepts (FC1) has answered to the question, who is involved in the offence Burglary Other Building or Attempt Theft of Vehicle. All the objects of extent in this topmost node are the offenders who have committed these offences. Revealing the temporal information down the hierarchy concept FC6 reflects the answer to the specific query: Who committed offence Burglary Other Building in February. Likewise the formal concept (FC12) unfolds further information to answer the query who committed offence Burglary Other Building in February on Saturday morning. A formal Context like this is also capable to answer the location based queries, for example who committed offence Burglary Other Building in February on Saturday morning on King George Crescent? the answer being Macek Sanzone as the extent of the concept.

### B. Offender Network

Formal Concepts based lattice may also reveal such association between the offenders thus indicating possible crime networks. A criminal network is established when an offender commits one or more crimes with another offender(s) (1st degree of freedom) and those offender(s) themselves commit crime(s) with other offender(s) (2nd degree of freedom) [18]. We created another formal concept lattice keeping the offenders as objects. However, this time taking crime reports represented through crime reference number along with the type of the offence and spatial and temporal information about the committed crimes as the attributes of the object.

Table II Temporal and Spatial Profile of Offenders

<table>
<thead>
<tr>
<th>Crime Type</th>
<th>Offender</th>
<th>Time</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>Burglary Other Building</td>
<td>Clouser, Horethm, Khong</td>
<td>6 a.m. to 12 a.m.</td>
<td>Stonnal Road</td>
</tr>
<tr>
<td>Attempt Theft of Vehicle</td>
<td>Macek Sanzone</td>
<td>6 p.m. to 12 p.m.</td>
<td>King George Crescent</td>
</tr>
</tbody>
</table>

Figure 1. Formal Concept Lattice for WHO based Queries
A subset of crime data burglary of the other building and burglary dwelling for a year is used for this. The generated lattice consisted of 29 concepts, revealing offender network information based on the crime reference number along with other related attributes about the associated offenders. For example, Figure 2(a) shows criminal network through highlighted nodes of offender Cheng. The offender Cheng is connected to offender ENIX with the crime reference 2561253 both committed “Burglary Other Buildings” at Birmingham Street in May on Tuesday afternoon. Cheng, however, is also associated with another offender Yasin in committing the similar offence on Birmingham Street in May on Sunday afternoon through (crime reference 2561040). The lattice also reveals network association between three offender Cheng, Yasin and Haven for committing the similar offence (crime reference 260192) same year together at Wall Street in July on Monday evening. The faded nodes are for non-associated offenders not included in Cheng Network. Another network that is visible in the lattice is shown in Figure 2(b) is of the offender Teas who is associated with Turtoo in committing “Burglary Other Buildings” on George Street in December on Thursday afternoon through Crime_Ref (2729327). Offenders association for another offence “Burglary Dwelling” is also visible: the offender Capasso is associated with offender Kund in committing the offence on Petersham Road in Decon on Monday night. Other networks are also visible.

C. Crime Hot Spots

Costa [19] has defined a Crime Hot Spot (CHS) as a geographical area having a higher incidence of crimes than their neighboring areas and represent areas where people are more likely to be victimized. The analysis of CHS is helpful in distribution of resources such as policemen and patrol cars and for defining strategies for the prosecution of crime. These CHSs are characterized by the time of the occurrence offence, total number of offences in the area and spatial information of the area and hence temporal-spatial analysis of the crime is the key to analyse the CHS. We used FCA to identify the CHSs using spatial information of offences as object and different types of crimes in different periods of a day as attributes to identify the CHS we took three different crimes "Burglary Other Building (BOB), 'Burglary Dwelling, 'Theft From Shop OR Stall"” for a calendar year. The generated cross table is shown in Table 3.0 followed by the Formal Concept Lattice Figure 3. The CHS FCA lattice represents a hierarchy of questions starting from general to specific for example, show the crime hotspot for burglary of the other building and theft from shop or stall for day time intervals. We observe that no offences occurred during night or evening, they occurred either morning or afternoon. Afternoon is found to be the most active period of the day for both the offences however they occurred on separate days. We can also observe from the lattice that the crime hot spots are different for the three crimes. We observe that there are no reported thefts from shop or stall.

VI. ASSOCIATION RULES

The A Priori algorithm is a classic and probably the most basic algorithm employed in data mining for finding association rules, through identifying frequent item-sets appearing together in the data. An association rule takes the following form: X → Y, where X (conditions) and Y (implication) are sets of atomic propositions. The association rules are used in a wide range of applications, due to their comprehensive format. For example, Amazon.com, use association mining to recommend users the items based on the current item a user is browsing or buying. Likewise Google auto-complete feature searches frequently associated words the user types after that particular word. We employed frequent itemset mining through the A Priori algorithm to find the association rules between the offenders and their associated offences along with spatial and temporal information.
Confidence (X → Y) is the number of examples satisfying both X and Y, divided by the number of examples satisfying only the condition X. The generated set of association rules are presented in Table III along with support and confidence levels.

<table>
<thead>
<tr>
<th>Antecedent</th>
<th>Consequence</th>
<th>Support</th>
<th>Confidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 {BURGLARY OTHER BUILDING}</td>
<td>{Morning}</td>
<td>57.14%</td>
<td>72.72%</td>
</tr>
<tr>
<td>2 {ATTEMPT THEFT OF MOTOR VEHICLE}</td>
<td>{LICHFIELD ROAD, Night, Sep, Thu}</td>
<td>21.42%</td>
<td>100.0%</td>
</tr>
<tr>
<td>3 {STONNALL ROAD}</td>
<td>{BURGLARY OTHER BUILDING, Feb, Night, Sun}</td>
<td>21.42%</td>
<td>100.0%</td>
</tr>
<tr>
<td>4 {Feb}</td>
<td>{BURGLARY OTHER BUILDING}</td>
<td>42.85%</td>
<td>100.0%</td>
</tr>
<tr>
<td>5 {Feb, Morning}</td>
<td>{BURGLARY OTHER BUILDING, Sat}</td>
<td>21.42%</td>
<td>100.0%</td>
</tr>
<tr>
<td>6 {Jun}</td>
<td>{Afternoon, THEFT FROM SHOP OR STALL, Wed}</td>
<td>21.76%</td>
<td>66.66%</td>
</tr>
<tr>
<td>7 {Apr}</td>
<td>{Morning, SAGOES, SHACKLE, Sun, THEFT FROM SHOP OR STALL}</td>
<td>31.76%</td>
<td>66.66%</td>
</tr>
<tr>
<td>8 {CHENG}</td>
<td>{BURGLARY OTHER BUILDING, YASIN}</td>
<td>11.76%</td>
<td>100.0%</td>
</tr>
<tr>
<td>9 {SAGOES}</td>
<td>{Apr, Morning, SHACKLE, Sun, THEFT FROM SHOP OR STALL}</td>
<td>11.76%</td>
<td>100.0%</td>
</tr>
<tr>
<td>10 {THEFT FROM SHOP OR STALL, Tue}</td>
<td>{Dec, HALLOWELL, PAILTHORPE}</td>
<td>11.76%</td>
<td>66.66%</td>
</tr>
</tbody>
</table>

VII. CONCLUSION

In this work we have presented the idea of associative search based on 5WH questions and employed FCA theory to implement it. We found while FCA can be used to reveal the association such as offender network, CHSs, however the application of FCA can generate thousands of questions, making browsing the hierarchy cumbersome. We also plan to study a set of measures, inspired both from ontology and FCA, to identify the questions that are more likely to be close to the ones of interest to the user.

ACKNOWLEDGMENTS

The research leading to the results reported here has received funding from the European Union Seventh Framework Programme through Project VALCRI, European Commission Grant Agreement N° FP7-IP-608142, awarded to Middlesex University and partners. The data used in the examples were anonymised from actual data produced to support the project.

REFERENCES


