Title: Literature Review and Results for Motivation

Author: Blake Conrad

Email: [bmconrad@iupui.edu](mailto:bmconrad@iupui.edu)

Project Advisor: Dr. Murat Dundar

|  |
| --- |
| Literature Review |

For every problem, there have been some solutions, exact or relevant, as systems (commercial) or in research papers and manuals. Study them, find pros, cons, or constraints ect, and describe them.

* What is K-means?
* What are its strengths?
* What are its weaknesses?

|  |
| --- |
| Results for Motivation |

Justify why you are choosing a specific solution, given other existing solutions. The justification has to be experimental, theoretical (if Math is involved), or practical (e.g., application constraints).

Your chosen solution has to tackle specific challenges that others do not tackle.

* How does it stand up to Crime Data Pattern Detection? Has this particular problem been studied?
* Make mention of the ranking system to determine favorable crime behavior over time.

Use the following as a way to site my sources in-text and on my citation page.

|  |
| --- |
| Even for Computer Scientists with exposure to big data an undergraduate level, as studied byRobert Jones a professor at the University of Washington, there are gaps in education. As statedby Jones, “US academic institutions tend to lack courses in several topics required forprofessional software production. As a result, recent college graduates in any technical domainusually require mentoring and on-the-job training before they can be entrusted with significantassignments” (Jones 1995, 70)  Jones, C. Gaps in programming education (1995). Computer Volume: 28, Issue: 4, 70-71.doi:10.1109/2.375185. |

Use the following as a resource for motivation in the subject and **why it’s relevant**.

|  |
| --- |
| [H. Chen](http://ieeexplore.ieee.org/search/searchresult.jsp?searchWithin=%22Authors%22:.QT.H.%20Chen.QT.&newsearch=true); [W. Chung](http://ieeexplore.ieee.org/search/searchresult.jsp?searchWithin=%22Authors%22:.QT.W.%20Chung.QT.&newsearch=true); [J.J. Xu](http://ieeexplore.ieee.org/search/searchresult.jsp?searchWithin=%22Authors%22:.QT.J.J.%20Xu.QT.&newsearch=true); [G. Wang](http://ieeexplore.ieee.org/search/searchresult.jsp?searchWithin=%22Authors%22:.QT.G.%20Wang.QT.&newsearch=true); [Y. Qin](http://ieeexplore.ieee.org/search/searchresult.jsp?searchWithin=%22Authors%22:.QT.Y.%20Qin.QT.&newsearch=true); [M. Chau](http://ieeexplore.ieee.org/search/searchresult.jsp?searchWithin=%22Authors%22:.QT.M.%20Chau.QT.&newsearch=true)  Crime data mining: a general framework and some examples  **Page(s):**50 - 56  **Date of Publication:** 02 August 2004  **DOI:**[10.1109/MC.2004.1297301](https://doi.org/10.1109/MC.2004.1297301)  Hsinchun Chen, Wingyan Chung, Yi Qin, Michael Chau, Jennifer Jie Xu, Gang Wang, Rong Zheng, Homa Atabakhsh, "Crime Data Mining: A General Framework and Some Examples", IEEE Computer Society April 2004. |

* Why is data mining important for crime data?
  + Make mention of the author’s, their research & expertise, and their take on this issue from a general theoretical point of view.
* How can data mining be used to find effective patterns in the crime data?
* This is where I mention that my project has a twist that other research has not explicitly defined as a problem/solution approach.

Concern about national security has increased significantly since the terrorist attacks on 11 September 2001. The CIA, FBI, and other federal agencies are actively collecting domestic and foreign intelligence to prevent future attacks. These efforts have in turn motivated local authorities to more closely monitor criminal activities in their own jurisdictions.

Two examples given that are applicable to Data Mining Crime Patterns:

1. *For example, complex conspiracies are often difficult to unravel because information on suspects can be geographically diffuse and span long periods of time. (pg 50)*
2. *Detecting cybercrime can likewise be difficult because busy network traffic and frequent online transactions generate large amounts of data, only a small portion of which relates to illegal activities. (pg 50)*

Different Data Mining Techniques used to find crime patterns:

1. *Entity Extraction* (identifies particular pattern from data such as text, images, or audio materials)
2. *Clustering Techniques* (used to group items into classes with similar characteristics to maximize or minimize intraclass similarity)
3. *Association rule mining* (discovers frequently occurring item sets in a database and presents the patterns as rules)
4. *Sequential mining* (finds frequently occurring sequences of items over a set of transactions that occurred at different times)
5. *Deviation detection* (uses specific measures to study data that differes markedly from the rest of the data),
6. *Classification* (finds common properties among different crime entities and organizes them into predefined classes),
7. *String Ccomparator Techniques* (compares textual fields in pairs of database records and computes the similarity between the records)
8. *Social Network Analysis* (describes the roles of and interactions among nodes in a conceptual network)

These are all of the listed types of techniques generally used for this type of problem.

## Authors

[H. Chen](http://ieeexplore.ieee.org/search/searchresult.jsp?searchWithin=%22Authors%22:.QT.H.%20Chen.QT.&newsearch=true)

Arizona Univ., Tucson, AZ, USA

Hsinchun Chen is the McClelland Professor of Management Information Systems at the University of Arizona's Eller College of Business and Administration. His research interests include medical informatics, homeland security. semantic retrieval, search algorithms, knowledge management, and Web computing. Chen received a PhD in information systems from New York University. Contact him at hchen@eller.arizona.edu.

[W. Chung](http://ieeexplore.ieee.org/search/searchresult.jsp?searchWithin=%22Authors%22:.QT.W.%20Chung.QT.&newsearch=true)

Arizona Univ., Tucson, AZ, USA

Wingyan Chung is a doctoral candidate in the Department of Management Information Systems at the University of Arizona, where he is a research associate in the Artificial Intelligence Lab. His research interests include knowledge management, knowledge discovery on the Web, text mining, security informatics, and human-computer interaction. Chung received an MS in information and technology management from the Chinese University of Hong Kong. He is a member of the IEEE Computer Society, the ACM, and the Association for Information Systems. Contact him at wchung@eller.arizona.edu.

[J.J. Xu](http://ieeexplore.ieee.org/search/searchresult.jsp?searchWithin=%22Authors%22:.QT.J.J.%20Xu.QT.&newsearch=true)

Arizona Univ., Tucson, AZ, USA

Jennifer Jie Xu is a doctoral candidate in the Department of Management Information Systems at the University of Arizona, where she is a member of the Artificial Intelligence Lab. Her research interests include knowledge management, social network analysis, computer-mediated communication, and information visualization. Xu received an MS in computer science from the University of Mississippi. She is a member of the IEEE Computer Society. Contact her at jxu@eller.arizona.edu.

[G. Wang](http://ieeexplore.ieee.org/search/searchresult.jsp?searchWithin=%22Authors%22:.QT.G.%20Wang.QT.&newsearch=true)

Arizona Univ., Tucson, AZ, USA

Gang Wang is a doctoral student in the Department of Management Information Systems at the University of Arizona. His research interests include deception detection, data mining, Web mining, and knowledge discovery. Wang received an MS in industrial engineering from Louisiana State University. Contact him at gang@eller.arizona.edu.

[Y. Qin](http://ieeexplore.ieee.org/search/searchresult.jsp?searchWithin=%22Authors%22:.QT.Y.%20Qin.QT.&newsearch=true)

Arizona Univ., Tucson, AZ, USA

Yi Qin was a member of the Artificial Intelligence Lab at the University of Arizona. Her research interests include medical informatics and cyber-crime analysis. Qin received an MS in management information systems from the University of Arizona. Contact her at yiqin@eller.arizona.edu.

[M. Chau](http://ieeexplore.ieee.org/search/searchresult.jsp?searchWithin=%22Authors%22:.QT.M.%20Chau.QT.&newsearch=true)

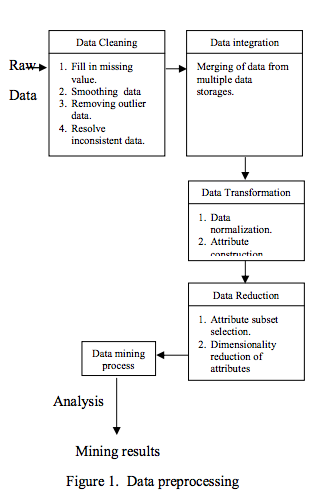
Michael Chau is a research assistant professor in the School of Business at the University of Hong Kong and was formerly a research associate at the University of Arizona's Artificial Intelligence Lab. His research interests include text mining, Web mining, digital libraries, knowledge management, and intelligent agents. Chau received a PhD in management information systems from the University of Arizona. He is a member of the IEEE Computer Society, the ACM, the Association for Information Systems, and the American Society for Information Science and Technology. Contact him at mchau@business.hku.hk.

Use the following as a **survey of other techniques used**:

|  |
| --- |
| <http://s3.amazonaws.com/academia.edu.documents/34277088/DEC_12_IJDMTA_011.pdf?AWSAccessKeyId=AKIAIWOWYYGZ2Y53UL3A&Expires=1486749343&Signature=hIDsO4s52YFDK8QoZfihFs7InGo%3D&response-content-disposition=inline%3B%20filename%3DSURVEY_OF_DATA_MINING_TECHNIQUES_ON_CRIM.pdf> |

Vol 01, Issue 02, December 2012 International Journal of Data Mining Techniques and Applications http://iirpublications.com ISSN: 2278-2419

The authors state the three most effective forms of clustering for the Crime Pattern Data Mining

* The partition clustering methods primarily classified into K-means, AK-mode and Expectation-Maximization algorithms. The partitioning method constructs ‘k’ partitions of the data from a given dataset of ‘n’ objects.
* The image on this page shows how their data preprocessing took place, and what techniques were used to format the data. Note: I did use the preprocessing transformation into centered data, as well as uncentered to determine the best K. I also used population as s crime statistic, as my project advisor stated, the population volume is critical in assessing the health of the city. Note: I did not use the data reduction technique here yet. This was not initially the goal of the project as agreed upon with the project advisor, however, if time permits I want to see if the PCA algorithm helps in determining K.
* Crime data is a sensitive domain where efficient clustering techniques play vital role for crime analysts and law-enforcers to precede the case in the investigation and help solving unsolved crimes faster.

Use the following as a resource for **motivating the pro’s and con’s of Kmeans directly on the subject**.

|  |
| --- |
| [Shyam Varan Nath](http://ieeexplore.ieee.org/search/searchresult.jsp?searchWithin=%22Authors%22:.QT.Shyam%20Varan%20Nath.QT.&newsearch=true)  Florida Atlantic University/Oracle Corporation, USA  <http://ieeexplore.ieee.org/document/4053200/>  Crime Pattern Detection Using Data Mining  **DOI:**[10.1109/WI-IATW.2006.55](https://doi.org/10.1109/WI-IATW.2006.55) |

* Why Kmeans?
  + It’s fast, it’s generally a good starting place to lean onto other algorithms from, sometimes is good enough for the mining that needs to be done.
* Objections?
  + If dimensionality reduction isn’t done, some would say that EM clustering will be more effective to take into account the variance of the Gaussian distributions and the probability of a data point falling into one of the Gaussian’s is sometimes more accurate than diagnosing a data point to a cluster explicitly.
* We used k-means clustering technique here, as it is one of the most widely used data mining clustering technique. Next, the most important part was to prepare the data for this analysis. The real crime data was obtained from a Sherriff's office, under non-disclosure agreements from the crime reporting system. The operational data was converted into denormalised data using the extraction and transformation. Then, some checks were run to look at the quality of data such as missing data, outliers and multiple abbreviations for same word such as blank, unknown, or unk all meant the same for missing age of the person.
* Thus clustering algorithms in data mining are equivalent to the task of identifying groups of records that are similar between themselves but different from the rest of the data. In our case some of these clusters will useful for identifying a crime spree committed by one or same group of suspects. Given this information, the next challenge is to find the variables providing the best clustering. These clusters will then be presented to the detectives to drill down using their domain expertise.
* We choose to use clustering technique over any supervised technique such as classification, since crimes vary in nature widely and crime database often contains several unsolved crimes. Therefore, classification technique that will rely on the existing and known solved crimes, will not give good predictive quality for future crimes. Also nature of crimes change over time, such as Internet based cyber crimes or crimes using cell-phones were uncommon not too long ago. Thus, in order to be able to detect newer and unknown patterns in future, clustering techniques work better.
* Successes: *Our modeling technique was able to identify the crime patterns from a large number of crimes making the job for crime detectives easier. (Pg. 22)*
* Shortcomings: *Some of the limitations of our study includes that crime pattern analysis can only help the detective, not replace them. (Pg. 22)*