i-JEN: Visual Interactive Malaysia Crime News Retrieval System

Nazlena Mohamad Ali[1], Masnizah Mohd[2], Hyowon Lee[3], Alan F. Smeaton[3], Fabio Crestani[4] and Shahrul Azman Mohd Noah[2]

[1]Institute of Visual Informatics, University Kebangsaan Malaysia, Malaysia
[2]Faculty of Information Science and Technology, University Kebangsaan Malaysia, Malaysia
[3]CLARITY: Centre for Sensor Web Technology, Dublin City University, Ireland
[4]Faculty of Informatics, University of Lugano, Switzerland
{nma, mas, samn}@ftsm.ukm.my, {hlee, asmeaton}@computing.dcu.ie, fabio.crestani@usi.ch

Abstract. Supporting crime news investigation involves a mechanism to help monitor the current and past status of criminal events. We believe this could be well facilitated by focusing on the user interfaces and the event crime model aspects. In this paper we discuss on a development of Visual Interactive Malaysia Crime News Retrieval System (i-JEN) and describe the approach, user studies and planned, the system architecture and future plan. Our main objectives are to construct crime-based event; investigate the use of crime-based event in improving the classification and clustering; develop an interactive crime news retrieval system; visualize crime news in an effective and interactive way; integrate them into a usable and robust system and evaluate the usability and system performance. The system will serve as a news monitoring system which aims to automatically organize, retrieve and present the crime news in such a way as to support an effective monitoring, searching, and browsing for the target users groups of general public, news analysts and policemen or crime investigators. The study will contribute to the better understanding of the crime data consumption in the Malaysian context as well as the developed system with the visualisation features to address crime data and the eventual goal of combating the crimes.

Keywords: interaction design, visualisation, content analysis, crime news

1 Introduction

Crime in Malaysia has been steadily increasing in the past few years. The national crime rates last year increased by 15.74% with every states reporting an increment in

2006[1]. The Index Crime statistic has proved that the crime situation in Malaysia is critical and newspaper reporting shows that the public continues to see crime as one of the most pressing problems in society [21]. In a layman's term Index Crime represents those offences that are regular and common in occurrence thus it can be later used to compare general crime situation between countries.

The journalists and the general public mainly depend on resources reported by the media production in investigating or monitoring crime topics. Therefore a huge amount of crime news needs to be organized in an effective way such as mining the crime news should provide information on the crime pattern and to discover new information. It would be good to have a crime news system that will be able to automatically track the crime topics and detect new news for a specific crime. This is beneficial to users such as the journalist in writing news or to the police in monitoring the crime news, as well as general public in their day-to-day news consumption.

This paper is organised as follows. Section 2 elaborates some related works on event crime model and crime data visual interfaces. It is followed in Section 3 by our approaches in designing and developing Malaysian crime news system which consist of technical possibilities and visual interaction design. Finally we conclude our approaches in designing the system and some future work in Section 4.

2 Related Works

This section discusses the related works on event crime model and crime data visualization.

2.1 Event crime model

The notion of "event" is highly associated with a crime. When crime news such as 'Sosilawati murder' was reported, we tend to investigate and to know on 'who killed her?', 'when did she die?', 'where was she killed?' 'what happened?'. It was observed that when hearing a crime, what we concern about most may be its participants, time and location it happen and probably the instruments used and goods involved. The questions on the *Who, Where and When* are more to the factoid or facts finding meanwhile the *What* question is more on the associative discovery or finding the chronology of the murder case.

Brown [5] has constructed a software framework for mining data in order to catch professional criminals. He thought that the analyst needs to get the details of data such as where and when the incidents occurred. Nath [17] outlined a new approach for crime pattern discovery. In his method, each record is composed of many attributes describing the crimes, such as date/time, location, outline, demographic and weapon.

---

[1] Malaysia Crime Watch (accessed May 2011)
http://malaysiacrimewatch.lokety.com/malaysia-crime-rate-up-15-per-cent-in-2006/

An event is identified by event triggers, and is associated with participants, time, location and others, and is a larger semantic unit compared with a concept. There is an intrinsic link between events. It is a new attempt to apply the semantic analysis technology of events to mine web crime information on the web.

Cunhua et al. [7] have explored cyber crime in Chinese web pages by event ontology construction. In particular, they define event ontology and demonstrate how it can be used to describe cyber crimes on the level of event, relation and event class. Promising techniques such as SVM-based text classification was employed in the implementation of their prototype system. Some researchers have proposed ideas of event-oriented ontology for processing events. Although Sánchez and Moreno [19] did not mention the notion of events, he actually applied an event triple model to assist the construction of ontology. Lin and Liang [12] presented the method of information retrieval based on event ontology. Han [10] constructed a character ontology model based on events, in which a character will relate to some special events, and events are attributes of characters. But his ontology is still limited to the character ontology.

In the context of Topic Detection and Tracking (TDT) which is an area aims to effectively retrieve and organise broadcast news (speech) and newswire stories (text) into groups of events, Nallapati et al. [16] have modelled the news topic based on event and their dependency. They named the process of recognizing events and identifying dependencies among them as *event threading*, an analogy to email threading that shows connections between related email messages. Although their corpus are the general news and not the crime news, we believed it would be interesting to model crime news topics by considering a relational structure of events interconnected by dependencies. Again, the crime news such as '*Sosilawati murder'* starts with the missing event, followed by kidnapping and finally the killing event. Therefore event dependency could support the associative discovery or finding the chronology of a crime case.

2.2 Crime Data Visualization

A number of works has been done in applying visualization techniques for crime data, mostly focusing on either or both of geographic map views with crime locations plotted and timeline views with temporal frequency of crime events over time. Oakland Crimespotting[2] provides an exploration tool for crime data. The system supports panning and zooming on the geographical map representation and filtering the crime type. The application also has a feature to dynamically browse both based on time and day of the crime and temporal form of data visualization. Fig. 1 shows the interface of Oakland Crimespotting.

[2] http://oakland.crimespotting.org

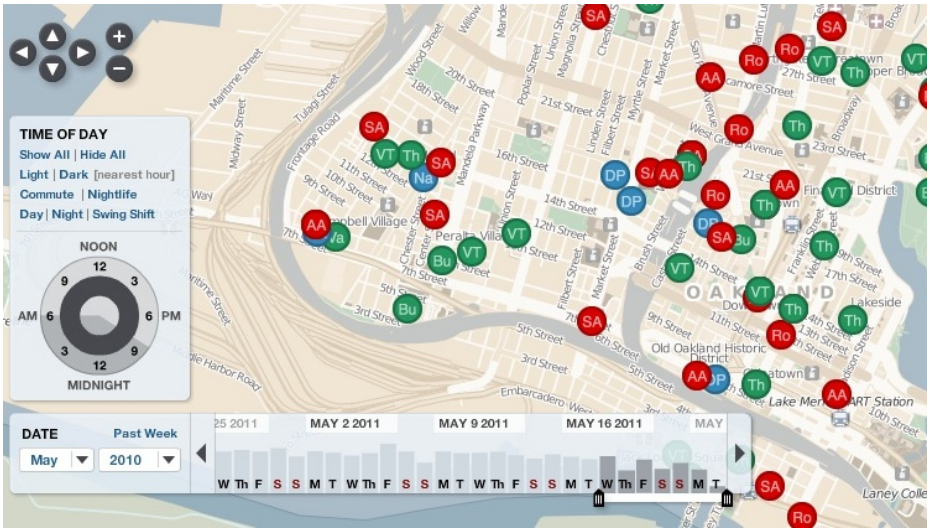

Fig. 1. Oakland Crimespotting interface

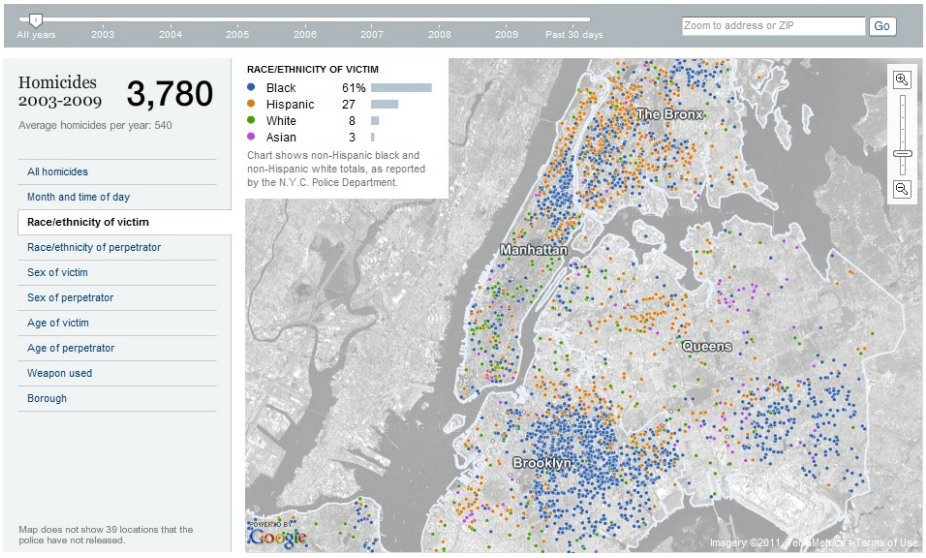

Fig. 2. New York Homicides Map interface

Another similar application is the New York Homicides Map[3] as shown in Fig. 2. The interface includes the geographical map representation and plots the locations of

[3] http://projects.nytimes.com/crime/homicides/map

the crimes, and supports searching and browsing of crimes by type and time, and also features temporal view of crime data.

WikiCrimes[4] is another related work in crime data visualization. It is a typical Web 2.0 application that allows users to access and register criminal events on the computer directly in a specific geographic location represented by a map. Fig. 3 shows the main screen of WikiCrimes, which offers a crime search function that enables users to view the registers of crimes, filtered by crime type and view statistics about the visualized area.

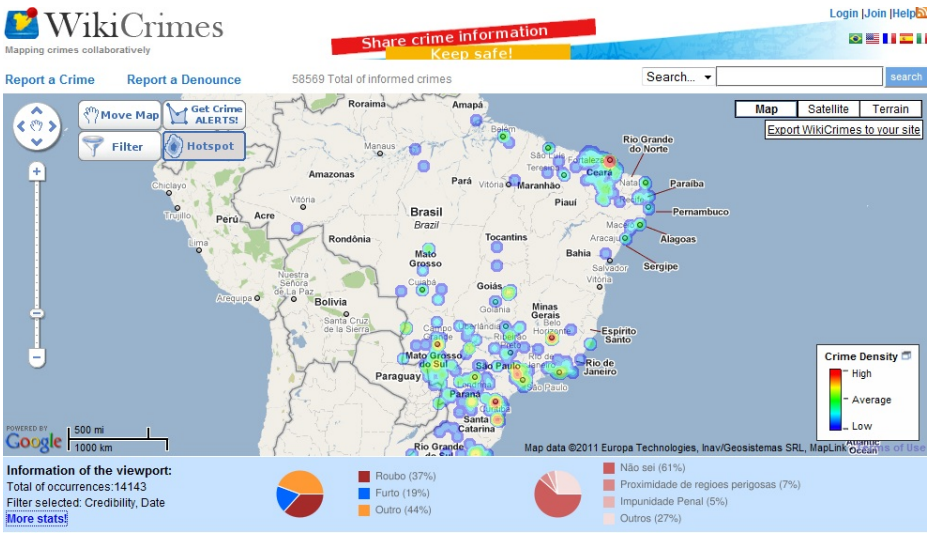

Fig. 3. WikiCrimes interface

Relatively a large number of crime-related visualisation systems are available using the similar concepts as the examples described above. While there are many visualisation ideas that have been explored in the past such as cone-tree, treemaps, Document Lens or Hyperbolic for different domain of data. Many specific techniques from these can be customised to be adopted to support the crime data visualisation. For example, dynamic query preview, zoomable, fish-eye lense, and small multiples are some of the techniques that can help visualising data while confirming the well-known information visualisation mantra "overview first, zoom and filter, then details on demand" [20]. It was also believed that visualizing a huge amount of document in cluster form helps user to understand news in a relatively fast and efficient manner [15],[22]. In addition, the use of advanced back-end processing of data (see Section

[4] http://www.wikicrimes.org

3.1) could result in more novel visualisation possibilities which currently we are not aware of.

3 An Interactive Malaysia Crime News Retrieval System (i-JEN)

i-JEN system is an interactive crime news monitoring system with the ability to track the crime news. It is a novel news system that applies the-state-of-the-art techniques in classification and clustering approach. This is a ground-breaking study which investigates Malaysia crime news where the focus is on the interactive crime data visualization, designed to help the general public, news analysts and the police to monitor the crime news. Fig.4 illustrates the conceptual framework of i-JEN.

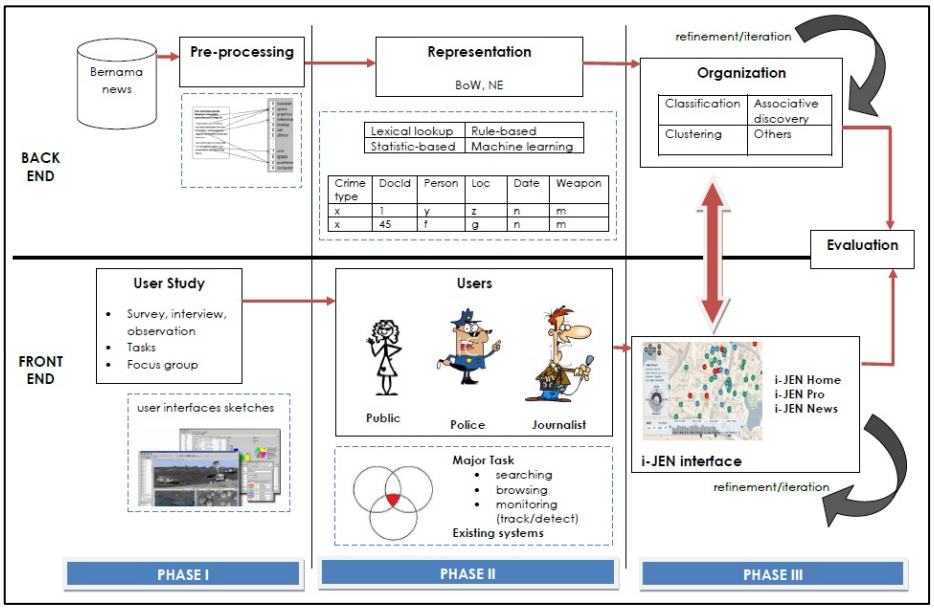

Fig. 4. i-JEN conceptual diagram

3.1 Technical Possibilities

There are three phases involved in i-JEN development approaches from the back end perspective as depicted in Fig. 4.

Phase I: Data Preparation and Preprocessing

This phase involve understanding the data provided by Bernama[5] news agency and preparing the crime news corpus and data preprocessing.

Characterizing and collecting the data - Understanding the characteristics of the crime-related information that needs to be organized and detected is important. This will ensure that the appropriate data will be collected and used to achieve good features in document representation. The data understanding activity starts with an initial data collection and proceeds with exploratory activities necessary to get familiar with the data, to identify data-quality problems, to gain an initial insight into the data or to identify interesting subsets, and to form hypotheses from the hidden information.

During this phase consideration is given to the quality of the data and how that will impact the results obtained. Consideration is also given to how we will access the data and address confidentiality and privacy issues. At this phase, considering additional information such as data from the police dataset or from judiciary, may be warranted. Table 1 shows some examples of the Malaysian high profile crime news topics that we have constructed into the solved case and unsolved case. Solved topic refers to the crime case that has gone through the prosecution and the sentence has been determined. On the other hand, unsolved topic refers to the ongoing crime case which is still under investigation process. Bache and Crestani have constructed the police dataset and they treated some of the solved case as the unsolved case [1].

Table 1: Construction of the high profile crime topics

Crime topic	Solved topic	Unsolved topic
Murder	Mona Fendy (**Maznah Ismail)** Noritta Samsuddin Canny Ong	Dato' Sosilawati Nurin Jazlin Jazimin
Kidnap	Dato' Sosilawati	Nurin Jazlin Jazimin Sharlinie Mohd Nashar

The researchers who are working in crime domain, used to build their own corpus by collecting data from multiple resources such as news portals and police databases [6]. Hence, the need for standard crime news corpus is on demand. This will be a challenge in i-JEN project since there is no Malaysia crime news corpus available and therefore, one of the contributions of the project is to collect, define and make accessible a set of Malaysia crime data for future researchers.

[5] http://www.bernama.com/bernama/v5/index.php

Data preprocessing - Data preprocessing mainly involves the elimination of the noisy data which may contribute to inaccurate results [23]. The first steps in data preprocessing phase is to remove the redundant stories and then the stopwords [3]. We believe that identifying a list of stopwords for crime domain will reflect significant improvement on feature selection process. Other techniques are also applied in this phase such as words stemming and tokenization [23].

Phase II: Data Representation

Data representation phase is important since the crime stories are represented in different formats. Researchers have modified and improved techniques in data representation such as finding new similarity distance measure, calculating similarity between objects and better data representation [4].

Each story in this phase is represented by a set of terms which is called *features*. Selecting the features from the indexed words is considered as a real challenge. This is because more than one feature or combinations of features are considered in this phase. Usually, the high score terms are selected as features, where the score of each term is calculated based on some predefined criterion such as TF-IDF weight [9]. This weight is a statistical measure used to evaluate how important a word is to a document in a collection or corpus.

Phase III: Data Organization

Data mining techniques have been used to extract useful information from a huge amount of unstructured data. Clustering analysis plays an important role in topic detection field [8]. The clustering techniques represent the process of segmenting a collection of various stories into subset where each subset groups the most similar documents together based on their similarity.

Clustering is widely used in TDT field, aiming to describe and construct the actual hierarchical of the news contents in simplifying the process of topic detection. In consequence, the clustering processes enhance the IR systems by speeding up the process of browsing and detecting the information needed [2].

3.2 Interaction Design and Visualization

While the accuracy and the speed of the processing and retrieval of relevant pieces of information are the key for the back-end of the system (top row in Fig. 4), the usability of the user-interaction and visualisation in terms of ease of use and the ability to reveal the patterns in the data are the key for the front-end of the system (bottom row in Fig. 4). Taking a strongly user-centred design approach in implementing the front-end element of the system, we start with our target user groups (general public, news analysts and the police) and their needs while being cognisant of the new technical possibilities that might arise from the back-end development.

Among the approaches to the user centred design is a Usability Engineering (UE). UE approach to system development is an iterative design process [18]. Nielsen recommended a Usability Engineering lifecycle as compared to a Waterfall model in building a system. His point of view is that linear progression in a development process from one set of specifications to another set will not succeed because most users cannot read specifications. Nielsen suggested development should be divided into three main stages, namely: Pre-design (i.e. field studies, usability test); Design (i.e. iterative design, prototyping); and Post-design stages (i.e. real-use data collections and feedback). Example of work that have implemented the UE throughout its system design and development is MovieBrowser2 with a user evaluation carried out with a film studies students in Dublin City University [13],[14]. By utilizing UE approaches, a system can be design and developed tailored to end users need and requirement. On the other hand, the technical possibilities for the back-end engine will follow a model define by Jesus Mena [11] in which the objective is to establish more efficient and reliable crime data detection and classification. This model defines the generic data mining processes which are: Understand the investigation's objective; Understand the data; Data preparation; Modeling; Evaluation and Deployment.

Our main goal in designing i-JEN interface for Malaysian crime news content is to enhance user interaction in information access and in exploitation of a user tasks. News information is inherently temporal while the instances of the information could take various types (i.e. text, video, images), elements (i.e. who, where, when), relation (i.e. cause and effect, chronology) and events category (i.e. murder, kidnap). This information can possibly be encoded in a variety of interactive visual display in which a user can exploit for different types of cognitive tasks such as retrieval, analysis, comparison, and summarization. In this work, we will adopt the user-centered approached in the system design and development in order to ensure the specific needs of the target users are accommodated. A number of issues that will be addressed in our initial user study are as follows:

- Visualization techniques - Our main consideration is to find out the most efficient ways of encoding crime data and a better way to represent clustering and classification. These will relate to the types of data or documents used, and will be applied by using some of the techniques mentioned in Section 2.2.
- User Interaction - Most of the applications regarding crime data interaction design facilitate the function of filtering, searching, exploration and browsing. Thus our application will highlight these tasks in its design.
- Cognitive tasks – The design of the visualization and user interaction will also consider ways to support the users' cognitive tasks such as identification, comprehension or abstraction of the crime data.
- Evaluation - The application will be evaluated in term of its effectiveness of the visualization components, usability and the system performance. An experiment will be conducted in two forms which is longitudinal study of the deployment system and also as a controlled lab experiment with a sample group of users.

Our approach in designing the application will be based on simplified version of UE process as illustrated in our conceptual diagram shown in Fig. 4. We divided the tasks into three phases. In the first phase, we will conduct a user requirement study on potential group of users in the three groups of public users, news editors/journalists and the policemen as mentioned earlier. We began our work with the identification of user needs through observations, extensive document reviews, in-depth interviews and the focus groups. It will involves qualitative data analysis. We will soon commence sketching and prototyping a system that incorporates some functional features that might be useful for each focus group of users. A number of designs of low-fidelity system prototype will be used to gauge and capture initial user's opinion and feedback. The initial sketches that come from technical stream will emphasise on the technical possibilities in order to engage our sample users in a more open-ended brainstorming and discussions rather than fixated on the current practice only: illustrating technical possibilities that our users had not thought of will help the users see their work in different lights, allowing novel features to be conceived. Our main objectives at the first phase are to understand how the end user carries out their task particularly in managing crime data and help them see how those tasks might be conducted in different ways.

At the second phase, we will manage and analyse the data that we gathered from the first phase. At this phase, user needs for each of the different groups will be identified and categorised. It will then map into an intersection of user needs and requirement. An identification of the similarities or differences between user needs will be used as guidelines in designing future visualization tools. Our main focus will be the identification of their major tasks and grouping them into the features for searching, browsing or monitoring (i.e. tracking and detecting) of crime data. An extensive exploration of the state-of-the-art visualization tools will also be carried out.

The outcome of the second phase will be used to design and developed the application and the system will have another iteration and refinement based on user evaluation and feedback. Integration with the back-end engine and data processing will be performed. These are the identified design framework that will be used in design and developing Malaysia crime data visualization tool.

4 Conclusion

In this paper we presented our conceptual framework, work done so far and the plans for designing and developing an interactive visual Malaysian crime news retrieval system. By strongly focusing on the specific target user groups in the context of Malaysia and thus adopting user-centred design approach to rigorously ascertain their needs and requirements, we believe that the i-JEN project should be very useful and practical for the community in various ways: the crime data collected and organised will be useful for continuing research and development in information retrieval and management for the crime-related corpus; understanding the specific needs and requirements of the currently practice in crime information consumption in Malaysia will be useful to help guiding and directing future research and policy-making in this area; tailoring and customising the technological tools to the Malaysian crime context

will help identify new possible ways to support the user tasks in crime data usage and will open various avenues for further investigation.

Once the fully-working i-JEN system becomes operational, we will conduct a series of extensive longitudinal experiments with real users in order to understand the various features of the system we built and to see the ways those features are put into practice.

Acknowledgement

The work was supported by the University Kebangsaan Malaysia Arus Perdana research grant (UKM-AP-ICT-21-2010).

References

1. Bache R. and Crestani F. An Approach to Indexing and Clustering News Stories Using Continuous Language Models, Natural Language Processing and Information Systems, pp. 109--116, 2010
2. Bouras and Tsogkas V. Assigning Web News to Clusters. Proceedings of Conference on Internet and Web Applications and Services, pp. 1--6, 2010a
3. Bouras and Tsogkas V. Improving text summarization using noun retrieval techniques. Springer, Lecture Notes in Computer Science, 2010b
4. Brants T., Chen F., and Farahat A. A system for new event detection. In Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval, pp. 330--337, 2003
5. Brown D.E. The regional crime analysis program (RECAP): A frame work for mining data to catch criminals. Proc. of the IEEE International Conference on Systems, Man, and Cybernetics, pp. 2848--2853, 1998
6. Chandra, M. Gupta, and M. Gupta. A multivariate time series clustering approach for crime trends prediction. Proceedings of International Conference on Systems, Man and Cybernetics,pp. 892--896, 2008
7. Cunhua Li, Yun H., and Zhaoman Z. An event ontology construction approach to web crime mining. Seventh International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), vol.5, pp.2441--2445, 2010
8. Dai X., Chen Q., Wang X., and Xu J. Online topic detection and tracking of financial news based on hierarchical clustering. In Proceeding of International Conference on Machine Learning and Cybernetics, pp. 3341--3346, 2010
9. Hao X. and Hu Y. Topic detection and tracking oriented to BBS. In Proceeding of International Conference on Computer, Mechatronics, Control and Electronic Engineering , pp. 154--157, 2010
10. Han Y. Reconstruction of People Information based on an Event Ontology. Proc. of International Conference on Natural Language Processing and Knowledge Engineering, pp.446--451, 2007
11. Jesus M. Investigative Data Mining for Security and Criminal Detection. B-H Publisher, 2003

12. Lin H. F. and Liang J. M.. Event-based Ontology design for retrieving digital archives on human religious self-help consulting. Proc. of 2005 IEEE International Conference on e-Technology, e-Commerce and e-Service, pp. 522--527, 2005
13. Mohamad Ali N. and Smeaton A.F. Are Visual Informatics Actually Useful in Practice: A Study in a Film Studies Context. Lecture Notes in Computer Science, In: Visual Informatics: Bridging Research and Practice, Volume 5857/2009, pp. 811--821, 2009
14. Mohamad Ali N., Smeaton A.F., Lee H. and Brereton P. Developing, Deploying and Assessing the Usage of a Movie Archive System. Lecture Notes in Computer Science, In: Human-Computer Interaction. Interacting in Various Application Domains, Volume 5613/2009, pp. 567--576, 2009
15. Mohd M., Crestani F., and Ruthven I., Design of an Interface for Interactive Topic Detection and Tracking, Lecture Notes in Artificial Intelligence 5822: 227-238, 2009.
16. Nallapati R., Feng A., Peng F., and Allan J. Event threading within news topics. In Proceedings of the thirteenth ACM international conference on Information and knowledge management (CIKM '04). ACM, pp. 446--453, 2004
17. Nath S.V. Crime pattern detection using data mining. Proc. of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, pp. 41--44, 2006
18. Nielsen J. Usability Engineering. Academic Press, Inc., 1993
19. Sánchez D. and Moreno A. A methodology for knowledge acquisition from the web. International Journal of Knowledge-Based and Intelligent Engineering Systems, pp. 453--475, 2006
20. Shneiderman B. and B000APPF64 Plaisant C. Designing the User Interface: Strategies for Effective Human-Computer Interaction (5th Edition). Addison Wesley. 2009
21. Sidhu A. S. The Rise of Crime in Malaysia: An academic and statistical analysis. Journal of the Kuala Lumpur Royal Malaysia Police College, No. 4, 2005
22. Spence R. Information Visualization: Design for Interaction (2nd Edition). Prentice Hall. 2007.
23. Yang Y., Pierce T., and Carbonell J. A study of retrospective and on-line event detection. In Proceedings of the 21st Annual International ACM SIGIR Conference, pp. 28--36,1998