ISI-KDD: SIGKDD Workshop on Intelligence and Security Informatics 2010

Christopher C. Yang
College of Information Science and Technology
Drexel University
3141 Chestnut Street
Philadelphia, PA 19104
Chris.Yang@drexel.edu

Hsinchun Chen
Department of Management Information Systems
University of Arizona
1130 E. Helen St.
Tucson, AZ 85721
hchen@eller.arizona.edu

Intelligence and Security Informatics (ISI) is concerned with the study of the development and use of advanced information technologies and systems for national, international, and societal security-related applications. ISI is an interdisciplinary research field involving academic researchers in information technologies, computer science, public policy, bioinformatics, medical informatics, and social and behavior studies as well as local, state, and federal law enforcement and intelligence experts, and information technology industry consultants and practitioners to support counterterrorism and homeland security missions of anticipation, interdiction, prevention, preparedness and response to terrorist acts. ISI events were started in 2003. The events included the annual IEEE International Conference on ISI, Pacific Asia Workshop on ISI (PAISI), and European ISI Conference. Since 2009, we started the ISI-KDD workshop. More details about the ISI events can be found at http://www.isiconference.org/.

ISI-KDD covered four major topics including (i) information sharing and data/text/Web mining, (ii) infrastructure protection and emergency responses, (iii) terrorism informatics, and (iv) enterprise risk management and information systems security. Besides the regular research papers, we also organized the ISI-KDD Challenge for participants to "find the more radical and infectious threads, members, postings, ideas and ideologies" in a Dark Web Portal of several complete multi-year extremist forums. Participants were free to develop novel computational techniques and algorithms, e.g., linguistic analysis, topic extraction, multilingual text parsing, sentiment analysis, social network analysis, time-series analysis, etc. for the challenge. Participants were requested to describe their methods and results in papers. A panel of terrorism study experts was invited to evaluate the validity and value of the computing results. In addition, we like to acknowledge NSF Computational Research Infrastructure (CRI) Program for supporting the Dar Web research, which make the Dar Web Portal possible for this Challenge.

It was our honor to have Dr. X. Sean Wang to give an invited talk at the beginning of the ISI-KDD 2010 workshop. Dr. X. Sean Wang was a program director at the National Science Foundation in the Division of Intelligent Information Systems and the Dorothean Professor in Computer Science in the College of Engieering and Mathematical Science at the University of Vermont. The title of his talk was "Privacy Research in the Context of Trustworthy Computing". In his talk, he presented that the loss of privacy was a serious concern to many people and organization given the society's permeating data gathering efforts and the highly connected world. This concern strained the adoption of computing technology to achieve its full potential for the benefit of the society, which was the goal of the Trustworthy Computing program at the National Science Foundation. He explained that privacy was the ability to control the release of information regarding persons and organizations of what they were, what they had, and what they did. Privacy was indeed an extension or refinement of confidentiality. Privacy was a control of who could access what but subjected to the society and users' rights/needs. A comprehensive solution of trustworthiness would be an appropriate combination of policy and technical solutions. Dr. Wang also gave examples of the recent work on privacy such as homomorphic encryption, longitude protocol for location monitoring, private information retrieval, secure computation, privacy-preserving data publication, etc.

The program committee accepted ten papers. Each of them had fifteen minutes of presentations.

In the paper titled "Two-stage Approach for Unbalanced Classification with Time-varying Decision Boundary: Application to Marine Container Inspection", Hoshino, Oldford and Zhu presented an approach to tackle the unbalanced classification problems which occurred when the Canada Border Services Agency (CBSA) was trying to identify the high-risk marine containers. In such problems, the optimal decision boundary might change over time. They proposed two models, namely smoothing model and regression model. The experiment result showed that both models outperformed the baseline mode and the regression model produced the best performance.

In the paper titled "Fuzzy Association Rule Mining for Community Crime Pattern Discovery", Buczak and Gifford studied the fuzzy rule mining to discover crime patterns at the community, state, and national levels. A relative support measure was proposed to prune the set of discovered rules and extract rare rules. A reduction of 95.2% was achieved in the experiment.

In the paper titled "The Structure and Content of Online Child Exploitation Networks", Frank, Westlake and Bouchard extracted the structure and features of four online child exploitation networks and conducted social network analysis to identifying key players. It was found that the presence of hardcore content was not the basis for linkages between websites but the blogs contained more hardcore content per page than websites.

In the paper titled "Real-Time Text Mining in Multilingual News for the Creation of a Pre-frontier Intelligence Picture", Piskorski, Atkinson, Belyaeva, Zavarella, Huttunen, and Yangarber presented a real-time event extract system to capture structured knowledge of border security-related events from online news. The proposed system was an effort for supporting Frontex – the European Agency for the Management of Operational Cooperation at the External Borders of the Member States of the European Union. The border security-related events included illegal migration incidents, cross-border criminal activities, and crisis events.

In the paper titled "Inference Control to Protect Sensitive Information in Text Documents", Cumby and Ghani presented the framework called Text Inference Control to ensure a user-specified level of privacy in text and maximize the information utility at the same time. They proposed the algorithms for batch inference control, pre-document inference control, and *k*-confusability. Both of the industry sector experiment and user evaluation experiment had shown promising results.

In the paper titled "Concept to Develop a Software-based Counter-Terrorism Campaign Decision Support Tool", Sinai presented a research note on the needs to develop a software tool kit to systematically display, track, operationalize and update the decisions and processes involved in addressing all the measures required to respond and resolve on-going terrorist-type insurgencies. Such tool kit could be employed for analytical and training purposes.

The last four papers were the ISI-KDD Challenge papers. These papers presented their effort in discovering knowledge from the Dark Web dataset. The Dark Web data set consisted of several multiyear extremist forums including seventeen forums in Arabic, seven forums in English, three forums in French, one forum in German, and one forum in Russian.

In the paper titled "Applying Interestingness Measures to Ansar Forum texts", Skillicorn presented an analysis of the documents from the Ansar aljihad forum and found that the content could be classified into two major categories. A model describing Salafist-jihadi content generated a single-factor ranking of postings which could be interpreted as the most radical postings. A model of deception created a multifactor ranking with low-deception postings identified with highly salafist-jihadi ones.

In the paper titled "Anomaly Detection in Extremist Web Forums Using a Dynamical Systems Approach", Kramer presented an approach of dynamic unsupervised anomaly detection. The objective was identifying the emerging threats in time-dependent and unlabelled data sets. The finite-time Lyapunov exponents were utilized to characterize the time evolution of the directed network structure and the distribution of text attributes in the forum messages. The identified anomalies were then related to the source data.

In the paper titled "Topic-Based Social Network Analysis for Virtual Communities of Interests in the Dark Web", L'Huillier, Alvarez, Rios, and Felipe Aguilera combined both social network analysis and text mining techniques to address the topic-based community key members extraction problem. They applied latent Dirichlet allocation to construct two topic-based social networks, one based on the thread creator and one based on the repliers of the overall forum. In the experiment using the Dark Web data, they identified 47 topics and 11 topics were manually discarded. The key members were extracted by HITS algorithm.

In the paper titled "An Analysis of User Influence Ranking Algorithms on Dark Web Forums", Yang, Tang, and Thuraisingham incorporated content similarity and response immediacy to measure the degree of influence between any two users in a social network site. The influence measure was utilized to construct a weighted social network. UserRank and weighted in-degree algorithms were proposed on the basis of PageRank and in-degree algorithms. An experiment was conducted to analyze the impact of the weights and basis algorithms on the user influence ranking using the Dark Web data set. It was found that the weights made a substantial difference on the ranking result but the basic algorithms did not made a substantial impact.

We are delighted of the success of the ISI-KDD workshop. We hope to continue the effort of this workshop in the future KDD conferences and welcome any comments, suggestions, or interests.