Applying Data Analytics to Development of the Web-Based Information Security Career Systems

Ming Wang, Hong Liu, and Drew Hwang

***Abstract*—This paper proposes a case to build a Web-based information/cyber security (ICS) career system for smart education using Web data analytical technology. The proposed system will automatically monitor and extract the new ICS job openings as well as the specific knowledge, skills, certificates or degrees required for the job openings from a variety of career Web sites, and classify these jobs by location, job title, skill, certificate, and degree requirement. This Web ICS career analytics system will serve the immediate needs of students and employers in the ICS field and will have great impacts on ICS job placement, curriculum development and industrial recruitment and work as a bridge to connect faculty and students in academia to prospective ICS employers and to improve ICS career education. The paper innovatively applies Web data analytical technology (data extraction, data mining, machine learning, data analytics and visualization) to ISC career development. The proposed case will be implemented with open source data analytics software tools GATE [1] and Weka [2].**

***Index Terms*—Data analytics, information security education, information security career, data mining**

I. Introduction

The field of information/cyber security (ICS) has grown rapidly in recent years with many career choices available: securing network(s) and allied infrastructure, securing applications and databases, security testing, information systems auditing, business continuity planning, cyber security and digital forensics. The abundant ICS career opportunities attract many students and even related industrial professionals to update their knowledge for midlife career change. To help college graduates and midlife career change people to get jobs in the dynamically grown ICS job market, the authors propose a case to develop this ICS career system using Web data analytical technology.

The idea of developing a smart ICS career Web system is original and novel. The Web-based ICS career system case is built with the authors' data analytics expertise, ICS knowledge, curriculum and career development experience. From a technology perspective, data analytics for the development of Web-based ICS career system goes through a number of technical processes including data streaming, machine learning, data mining, natural language processing and data visualization. The case will provide benefits to 1) students for their information security career development, 2) information security employers for their recruitment, and 3) faculty for their information security curriculum and course design. The goal of the case is to design and develop a Web-based data analytics system for ISC career development. The objectives of the case are:

- To connect ICS employers to college students who are interested in ICS career.
- To assist students who have taken ICS courses to find ICS internships and jobs.
- To guide students to select ICS related programs, courses, and build their career path
- To serve the need of people who want to make a mid-life career changes to ICS.
- To guide faculty in ICS curriculum reform and course design and update
- To help admission offices to reach out and student recruitment.

Manuscript received May 12, 2012; revised June 12, 2012.

Ming Wang is with Information Systems California State University, Los Angele (e-mail: ming.wang@calstatela.edu)

Hong Liu i is with Math and Computing, Embry-Riddle Aeronautical University (e-mail: liuho@erau.edu)

Drew Hwang is with Computer Information Systems, California State Polytechnic University (e-mail: dhwang@csupomona.edu)

II. Background

The authors' affiliated institutions of have integrated information security into their curriculum. California State University, Los Angeles has been working closely with California State Polytechnic University, Pomona Center for Information Assurance for the state of California. Both universities offer information security courses in their Computer Information Systems undergraduate program and graduate programs. Embry-Riddle Aeronautical University (ERAU) is also offers a Cyber Security Track under the Homeland Security Program and the university is strategically expanding the program to meet the growing need of security professionals regarding to aerospace and airport security issues.

Although the number of jobs in the field of Information/Cyber Security (ICS) has grown rapidly in recent years, our students who have taken some ICS courses still have hard time to find related ICS jobs. Similarly, ICS employers claim that they cannot find appropriately qualified applicants. The problem contributes to this disconnect between academia and ICS employers is one of the main reasons that has inspired the authors to develop the Web-based ICS Career Systems. Launching the smart ICS career system on the Internet not only serves the immediate needs of ICS students and educators in our institutions, but also meets the needs of ICS students and educators in other institutions as well as the needs of ICS employers. Furthermore classification and analysis of ICS job data will help ICS educators to adapt their courses and their curriculum to the industrial trends.

A. *Existing Technology*

Over the past decade, the Internet has created new

DOI: 10.7763/IJIEE.2012.V2.212

channels and enormous opportunities for organizations to post job openings [3]. The Internet opens great opportunities for employers and job applicants, but job openings posted on too many Web sites caused information overflow for job applicants. ICS jobs are mixed with other jobs on various career Web sites as well as some organizational Web sites.

B. *Proposed Technology*

In response the above problem, the proposed web-based system will monitor and extract ICS job openings from a variety of career Web sites as well as organization Web sites. The system is built with the open source data analytics software tools GATE [1] and Weka [2]. GATE is developed by University of Sheffield, UK. It is a framework and graphic development environment for natural language processing (NPL). Weka is developed by University of Waikato in New Zealand. GATE incorporates the widely adapted information extraction (IE) system Annie [4] and data mining system Weka [2]. In addition, it supports deep semantic analysis, machine learning, measuring, evaluating and benchmarking.

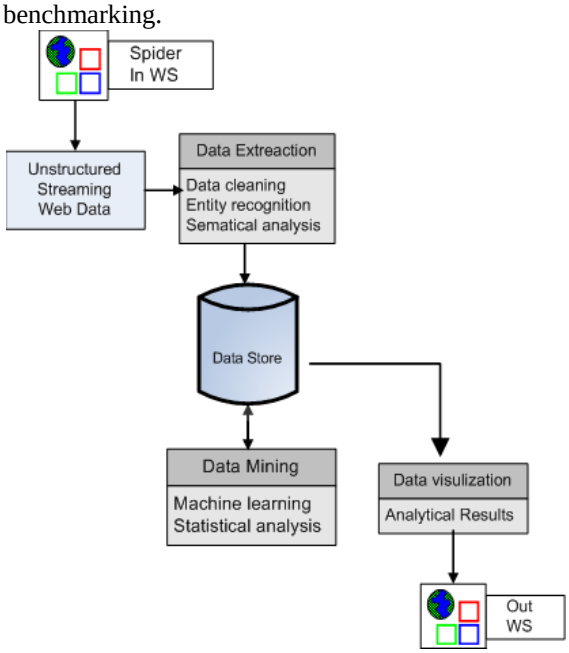

Fig. 1. Smart ICS career system overview

III. Case Description

The data analytics application for the smart ICS career Web site is built with GATE and Weka. On the top of the core functions GATE includes components for parsers, tagging, information retrieval, information extraction components. We choose to develop our smart ICS career Web site with GATE technology based on the three factors. Firstly, GATE CREOLE (collection of reusable of langrage engineering) packages all the resources for NLP and IE functions that we need such as JAR files, plus some XML configuration data. Secondly, GATE allows resource implementation and langrage resource persistent data to be distributed over the Web, and uses Java annotations and XML for configuration of resources. Thirdly, GATE documents, corpora and annotations can be stored in a variety of databases and be visualized via the development environment, and accessed at code level via the framework. These three factors really make our job relatively easy for the scope of our project. As a result, our Web-based data analytics application mainly provides a GUI, then controls the data access and configures the selected CREOLE plug-in resources. The classified job information will be automatically deployed on the internet. The smart ICS career Web site can be easily accessed and practically utilized by job applicants. No programming skills are needed from the end users.

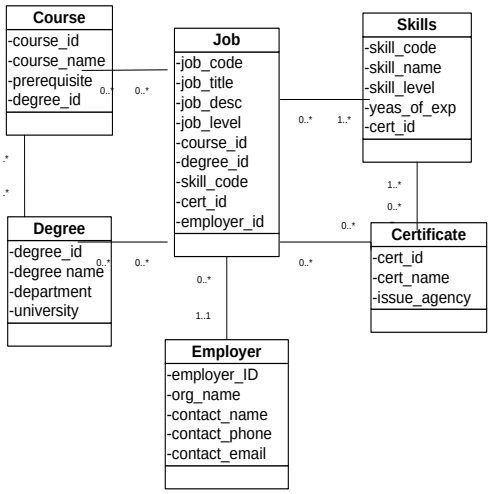

Fig. 2. Data model of smart ICS career web system

A. *Architecture Design*

The smart ICS career Web data analytical system overview is shown in Figure 1. Its main functions include Web services, information retrieval, information extraction, data store, data mining, and information visualization. The unstructured input documents are in HTML, XML, or RTF formats that are streamed down from Dice, CareerBuilder, Monster.com or Homeland security Web site via Internet search engine. We will use Annie [4] at the start of the learning phase. It performs text tokenization, sentence splitting, part of speech tagging, gazetteer lookup, and entity recognition. The GATE application then induces rules for job related feature extraction and summarization. The extracted features include name of employers, the job titles, years of working experiences; minimal education requirements, core skills, computer languages as shown in the class diagram in Figure 2. We plan to explore the three widely-used learning algorithms available in GATE PR (Processing Resources), Naïve Bayes, KNN and the C4.5 decision tree algorithm and select the most effective for the GATE application. We will use Weka [2] for job classification and computing the matching score based on the inputs about the qualification of the job seekers. The training and testing corpora for each text processing and classification tasks are manually selected. The precision, recall and F1 rate of BDM (Balanced Distance Metric) based IAA (Inter-Annotator Agreement) for each processing resource will calculated by the GATE Annotation Diff and Benchmarking Tools. The persistent data output of the application will be XML files that contain job records in the form of the data model shown in Figure 2. When the end user, e.g. a job seeker fills in the resume template in the web interface, the system first presents the matched jobs in spiral

circles and uses the vicinity to present the matching score intuitively. A variety of related analyses will be presented when users click the job item in the spiral circle.

C. *Implementation*

The project implementation includes three phases. In Phase One, we develop a GATE application for two purposes. One purpose is to retrieve documents from Internet to data store, and the other purpose is to extract standard data records from in data store. The data source is a large volume of unstructured documents. In phase two, we process data records and restore the records in the data store. The processed data records are saved as structured XML files and each record follows a template including all fields of the data attribute as described in the data model in Fig. 2. In phase three, we will build a Web application including two subcomponents. A data mining component will call Weka data mining library to classify and rank jobs when an end user input a job search request. A visual component will present the search results on the smart ICS career Web site.

D. *Data Model*

The application mainly uses three types of the data models. The first type is the retrieved job poster files. Most of the input files for job posters such as the those posted in Dice, which are a semi-structured file start with Job ID, follows Job tile (e.g. Information System Security Engineer), position description with a list of requirements containing technological keywords. Many job posters are unstructured documents described the job requirements in several paragraphs without particular order. Because the large volume of documents and the numerous features in each document, manually annotated learning and testing corpora for IE are too costly. Ontology driven markup and semi-supervised learning techniques [5]-[7] may be applied to improve data mining efficiency.

The common features are the keywords that IE can recognize such as Senior, Information Technology, Internet, Design, Graphic Design, Back End, End User, Software, Navigation, URL, Traffic, VP, IT, Network, M.S, Bachelor, etc. The second type of data is the input resumes of the job seekers (end users). It is a structured template with some item filled and some skipped. It includes fields such as degree, major, core computer courses, years of experiences and a few keywords to describe skills. The third data files are the persistent data output of the application after all the text engineering processing of our GATE application. It will be XML files that contain job records in the form of the data model shown in Fig. 2.

The job records are classified into clusters according some distinguishable technology features based on the ontology (taxonomy) of ICS security jobs. The machine learning algorithms for matching and scoring jobs with regard to a particular resume are based on the distance matrix between key features of the job records and resume records. At this stage, the data volume is not a major concern provided that it is large enough to annotate the training and testing data sets. Since we use streaming live data, it will have large volume in the future.

E. *Development Framework*

The development framework will include the following six phases

F. *Valuation and Dissemination*

The content of the project will be evaluated during the annual National Science Foundation (NSF) Curriculum Development in Security and Information Assurance (CDSIA) workshop that is organized annually by TRUST at San Jose State University. TRUST is the Team for Research in Ubiquitous Secure Technology, a NSF Science and Technology Center for development of security technology that will radically transform the ability of organizations (software vendors, operators, local and federal agencies) to design, build, and operate trustworthy information systems for our critical infrastructure. The feedback from ICS faculty at CDSIA will be collected at the end of the CDSIA workshop. Project evaluation forms will be given out during the presentation. In addition, the authors plan to conduct the following activities.

- To disseminate the discovery to the departments and admission/career offices of their affiliated universities.
- To survey ICS faculty via the TRUST and CDSIA e-mail lists three months after the project is deployed.
- To utilize Google Analytics [8] to record the number of visits, unique visitors, hits, and visiting time. The results will be displayed on the graphics and dynamic dashboard on the author's Google Site.

TABLE I: SIX PHASES OF THE DEVELOPMENT FRAMEWORK

Phase	Activities	Outcomes
I	Project planning, system analysis and design and Web Interface design	Data modeling, project schedule Web design storyboard
II	Identify, extract and classify key words Upload data manually	ICS database, Decision tree and Dataflow diagram
III	Analyze data , Configure the Web server Do data mining in the ICS data warehouse	ICS data warehouse, Project prototype, and Alive Web server
IV	Automatic data extraction and visualization features	Launch the complete project to the web site
V	Test and fix defects of Tune and revise the project	Release alpha, beta and final version of the project
VI	Evaluation and Dissemination	Assessment data

IV. SUMMARY

The Web-based ICS career development system is proposed to meet the challenge of the dynamic ICS job market and existing ICS courses and curriculum This Web analytical system not only can support job placement in the

ICS field, but also can provide support to the curriculum reform of the computing degree programs relevant to ICS field. The stored job requirement data can be utilized by the ICS educators to discover the ICS career trends. The current computing curricula are relative static and can hardly match the dynamics change of ICS job market. Academia needs to adapt such a change and update the ICS curriculum in the computing degree programs.

The proposed system can support data analytics and data mining of both static and streaming data from a variety of data sources of variable volume [9], [10]. It can handle increasing data volumes on the career Web site and on the ICS career Web site. The proposed Web data analytical system for ICS career is scalable. GATE [1] technology can be applied to build the system to allow diverse applications as plug-in components. The development framework and prototype can be extended to applications for other jobs and emerging degree programs such as healthcare information technology, aviation security and bioinformatics.

REFERENCES

[1] GATE, The University of Sheffield, 1995-2010, general architecture for text engineer [Online]. Available: http://GATE.ac.uk/

[2] W. Ka, "Machine Learning Group at University of Waikato," Data Mining Software in Java [Online]. Available: http://www.cs.waikato.ac.nz/~ml/weka/

[3] R. L. Melville, P. Perlich, and V. Sindhwani, et al. "Social Media Analytic," ***S****OR-MS Today. Linthicum*, 2010.

[4] Annie, Advanced Knowledge and Technology, Open Source Information Extraction fact-file [Online]. Available: http://www.aktors.org/technologies/annie/.

[5] M. V. Vera, E. Motta, J. Domingue, and et el, "Ontology Driven Semi-Automatic and Automatic Support for Semantic Markup," KMI [Online]. Available: http://projects.kmi.open.ac.uk/akt/publication-pdf/vargas-saakm02.pdf.

[6] M. Kavalec and V. Svatek, "Information Extraction and Ontology Learning Guided by Web," Cite Seer [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.18.6823.

[7] M. Chang, L. Ratinov, and D. Roth, "Use constraints to guide semi-supervised learning," ACL'07; ICML'08, Long'10, [Online]. Available: http://acl.ldc.upenn.edu/P/P07/P07-1036.pdf .

[8] B. Plaza, "Monitoring web traffic source effectiveness with Google Analytics An experiment with time series," in *Aslib Proceedings***,** Bradford, vol. 61, no. 5, pp. 474, 2009.

[9] E. Ikonomovska, S. Loskovska, and F. Gjorgjevik, "A Survey of Stream Data Mining," Cite Seer [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.65.8681.

[10] R. J. Mooney, and U. Y. Nahm, "Text Mining with Information Extraction, Multilingualism and Electronic Language Management*"* in *Proceedings of the 4th International MIDP Colloquium*, pp.141-160.