NORTHERN ARIZONA UNIVERSITY
The W. A. Franke College of Business

The Design and Evaluation of Multilingual Social Media Portal

Working Paper Series—11-09 | October 2011

Yan Dang[1]*, Yulei Zhang[1], Paul Jen-Hwa Hu[2], Susan A. Brown[3], Yungchang Ku[4], and Hsinchun Chen[3]

[1] Computer Information Systems,
The W. A. Franke College of Business,
Northern Arizona University, Flagstaff AZ, 86011
E-mails: {yan.dang, yulei.zhang}@nau.edu,

[2] Accounting and Information Systems,
David Eccles School of Business,
University of Utah, Salt Lake City, UT 84112
E-mail: paul.hu@business.utah.edu

[3] Department of Management Information Systems,
Eller College of Management
University of Arizona, Tucson AZ, 85721
E-mails: {suebrown, hchen}@eller.arizona.edu

[4] Computer Center,
Central Police University, Taiwan
E-mails: ycku1230@gmail.com

* Corresponding Author

The Design and Evaluation of Multilingual Social Media Portal

1. Introduction

Web 2.0 has created a large amount of user-generated content from online social media such as forums, blogs, social-networking sites, etc. (Chen, 2010). As a result, the volume of social media data has been growing exponentially. Such user-generated data contains valuable information about people's ideas and opinions toward products, services or social/political issues.

However, certain characteristics of user-generated content make it difficult to analyze this type of data. One characteristic is that most of such content is unstructured/semi-structured and stored across various data sources. Typically, there is no standard format when users post information on various social media sites such as forums, blogs, or social networking sites. Compared with other formal online data sources such as news websites and scientific dataset sources, such unstructured/semi-structured user-generated content first needs to be organized before further analysis. When people search information in social media sites, even for one specific topic or event, they will find more than one site (often a lot of sites) providing related information. Therefore, to use multiple data sources, another issue is raised - the data organizing procedure developed for one particular social media site typically cannot be applied to another.

Another characteristic is that much user-generated content is written in different languages. According to the latest statistics (updated in June 2010) of Internet users by language, 72.7% of Internet users are non-English speaking users (http://www.internetworldstats.com/stats7.htm). The top five groups of non-English speaking users are Chinese (24.2%), Spanish (7.8%), Japanese (4.7%), Portuguese (3.9%), and German (3.6%). Providing effective mechanisms to deal with multilingual data sources is critical for searching and analyzing social media data. Because of these two characteristics, traditional data analytical tools may not be able to access social media data or present it in a usable format (Kawamura, 2010). In addition, as the foundation for further data analytics, it is important to provide an automated, flexible way to access various social media sites in real-time, extract only the relevant content, and add structure to the data (Kawamura, 2010).

This study aims to develop an infrastructure for better social media data access. Specifically, an integrated framework of multilingual social media portal was developed. Three major functions were provided in the framework. They are Data Integration, Search Support, and Automatic Multilingual Translation Support. To evaluate the performance of the proposed portal framework, a user-oriented evaluation study was conducted to compare the performance of the portal with that of a benchmark system in terms of efficiency, effectiveness, system quality, ease of use, usefulness, user satisfaction, and intention to use.

The remainder of this paper is organized as follows. Section 2 discusses the data integration demand, multilingual issue, and spidering techniques related to social media data, as well as Web portals. Section 3 lists the research questions of this study. Section 4 describes the system development framework. Section 5 provides detailed information of the system evaluation study, including the prototype system, the benchmark system, research hypotheses, measurement variables, subjects and tasks, and data analysis and results. Section 6 summarizes this study and discusses future research directions.

2. Data Integration Demand, Multilingual Issue, Spidering Techniques, and Web Portals

The two-way communication enabled by Web 2.0 (O'Reilly, 2005) has fostered the exponential growth of user-generated content on the Web. Different from formal media, this new platform contains a vast amount of rich information about individuals' opinions and ideas. The followings are some examples.

- *Business Intelligence 2.0*: Companies try to identify ways to make profits by analyzing users' online activities, opinions and feedbacks (Chen, 2010; Kaplan & Haenlein, 2010).
- *Government 2.0*: Decision and policy makers try to establish real-time online interactions with citizens, businesses, and government agencies to further enhance the efficiency and effectiveness of the decision/policy making process (Chen, 2009; Palvia & Sharma, 2007).
- *Health 2.0*: Doctors, patients, and scientists utilize social media sites to provide personalized health care, collaborate, and promote health education (Hughes, Joshi, & Wareham, 2008).

2.1 Data Integration Demand of Social Media Data

Most user-generated content is unstructured/semi-structured. Recently, it has been reported that 95% of the 1.2 zettabytes of data in the digital universe is unstructured, 70% of which is user-generated content from social media (Roberts, 2011). Traditional data analytical tools may not be able to access the unstructured/semi-structured social media data or present it in a usable format (Kawamura, 2010). Therefore, integrating social media data has become the biggest challenge (Kawamura, 2010). An automated, flexible way to access various social media sites in real-time and to extract only the relevant content is in demand.

The goal of data integration of user-generated content is to enrich the unstructured/semi-structured data (e.g., online postings, comments, and conversations) by structuring them in a way that various types of information can be obtained - when it occurred, who said it, and what main points it conveyed (Kawamura, 2010). Although different social media sites often have different styles, after data integration, a consistent format can be obtained. Further data analytical tools can then be developed based on the consistent data format.

2.2 Multilingual Issue of User-Generated Content

When the Internet was first created in 1969 as DARPANET, it was dominated by English (Crystal, 2001). Nowadays, the World Wide Web contains information in more than 1,000 languages (Crystal, 2001). Since 2000, there have been more non-English-speaking users than English-speaking users on the Internet (Global Reach, 2004). According to the latest statistics (updated in June 2010) of Internet user by language, 72.7% of Internet users are non-English speaking users (http://www.internetworldstats.com/stats7.htm).

Since a lot of user-generated content has been written in different languages, an automatic translation support would be important to assist people's understanding of such multilingual content. Different methods have been explored to execute translation tasks (Abusalah, Tait, & Oakes, 2005; Zhou, Huang, & Chen, 2008; Zhou, Qin, Chen, & Nunamaker, 2005), such as dictionary-based approach, corpus-based approach, and machine translation-based approach.

In the dictionary-based approach, a bilingual dictionary is first constructed, and then a translation result is obtained by looking up a given term in the dictionary (Abusalah et al., 2005; Zhou et al., 2005). Depending on needs, the dictionary can be general or domain-specific. The corpus-based approach analyzes a large collection of documents to construct a statistical translation model (Abusalah et al., 2005; Zhou et al., 2005). A corpus is a repository consisting of textual documents, paragraphs and sentences written in one or more natural languages. The advantage of the corpus-based approach is that there is no need to rely on manually created bilingual dictionaries. Thus, this approach is more suitable for newly emerging domains where dictionaries are not available.

The machine translation-based approach uses existing machine translation techniques to provide automatic translation (Abusalah et al. 2005, Zhou et al. 2005). Compared with other approaches, it is relatively easy to apply and incorporate into an integrated system, is faster when performing the translation tasks especially for large texts, and no extra training is needed. Examples include Google Translation (http://translate.google.com/), Babel Fish (http://babelfish.altavista.com), and FreeTranslation.com (http://www.freetranslation.com). As one of the most popular machine translation tools, Google Translation provides translation functions for more than 80 languages.

2.3 Large Scale Data Spidering Techniques

In order to collect large-scale data from the Internet, spidering techniques are often utilized. Spiders are defined as software programs that retrieve Web pages and documents by traversing the World Wide Web following hypertext links (Cheong, 1996). There are six characteristics that are important for developing spiders, including accessibility, collection type, content richness, URL ordering features, URL ordering techniques, and collection update procedure (Fu, Abbasi, & Chen, 2010). Based on these characteristics, an effective spider needs to be able to deal with the registration of targeted sites (accessibility), collect and extract related information (collection type), filter out non-relative information (content richness), sort queued URLs (URL ordering features and techniques), and keep the collected information up-to-date (collection update procedure).

For social media portal, it is very important to collect the user-generated content in a timely manner. Therefore, the characteristic of collection update procedure is especially important. Two approaches are often used for collection update procedure, including periodic spidering and incremental spidering (Cho & Garcia-Molina, 2000). The periodic approach re-spiders the entire collection of targeted sites in a fixed time interval. The disadvantage of this approach is that when the size of the collection becomes huge, it needs to take a lot of time and resources to spider the entire collection. On the contrary, the incremental spidering approach focuses on collecting only the newly added content in the targeted sites. Therefore, the spidering process can be done within a much shorter time period.

2.4 Review of Web Portals

Web portals have been developed for various domains, such as business intelligence, health care, and emerging scientific areas (e.g. nanotechnology). However, most of them are in the traditional Web 1.0 era, utilizing Web 1.0 data sources such as Web pages and online scientific publications and documents. For example, Marshall et al. (2004) developed EBizPort, a Web portal for business intelligence, and conducted user evaluation to assess the performance of the system. Chung et al. (2004) developed a similar system, CBizPort, which supported business intelligence analyses in Chinese document environments. The system evaluation results indicated promising user satisfaction. In health care domain, Zhou et al. (2006) designed and implemented a Chinese medical portal (i.e., CMedPort), which allowed users to search for Web pages from local collections and meta-search engines, together with an encoding conversion between simplified and traditional Chinese to support cross-regional search and document summarization/ categorization. They conducted an experiment to evaluate the effectiveness, efficiency, error recovery, interface, and functionality of CMedPort, reporting that the use of CMedPort could result in significant improvements in users' search performance, compared with three benchmark regional search engines that included Sina, Yahoo! Hong Kong, and Openfind. The data sources for EBizPort, CBizPort, and CMedPort are all static Web pages.

Other Web portals have utilized online scientific publications, proceedings, documents, reports, and books. Examples include the Web of Science (http://scientific.thomson.com/products/wos/) by Thomson Scientific, the ACM Digital Library (http://portal.acm.org/dl.cfm) by the Association for Computing Machinery, the MEDLINE by

the National Library of Medicine (http://www.nlm.nih.gov/), the IEEE Computer Society Digital Library (http://www.computer.org/portal/site/csdl/index.jsp) by Institute of Electrical and Electronic Engineers, and Web-based systems maintained by leading patent offices to access patent documents (e.g., USPTO patent search system: http://patft.uspto.gov/). As to domain specific Web portals with analysis supports, for example, Dang et al. (2009) developed a Web portal, Arizona Literature Mapper, which allowed users to gain comprehensive understanding of current development status and emerging trend of research related to bioterrorism diseases by analyzing scientific publications from the major online database. The portal has practical importance for biodefense and national security. User evaluations of the portal performance were favorable. Nano Mapper (2011), a Web-based portal designed for nanotechnology domain, provided various types of search supports as well as analyses supports on data sources of nanotechnology-related patents and grant documents. Evaluation studies indicated the high effectiveness and efficiency of the system, as well as favorable user satisfaction toward using the system.

All the above systems used Web 1.0 data sources such as Web pages and online scientific publications and documents. None of them has utilized the Web 2.0 social media data sources, the content of which is user-generated with particular characteristics and demand of data integration of huge amount of unstructured data and real-time multilingual translation support.

3. Research Questions

As discussed in the above section, a systematic approach to integrate multilingual user-generated social media data is important and in demand. However, no previous research has developed an integrated framework to provide the access to multilingual, unstructured/semi-structured user-generated content from different social media data sources. This study aims to develop a Web portal for multilingual social media data and systematically evaluate the performance of the portal. The research questions that this study seeks to address are:

1. How to develop a multilingual social media portal that can effectively deal with data integration and the multilingual issue associated with the unstructured/semi-structured user-generated content?
2. How to systematically evaluate the performance of the portal by comparing it with alternative ways of searching information in multilingual user-generated content?

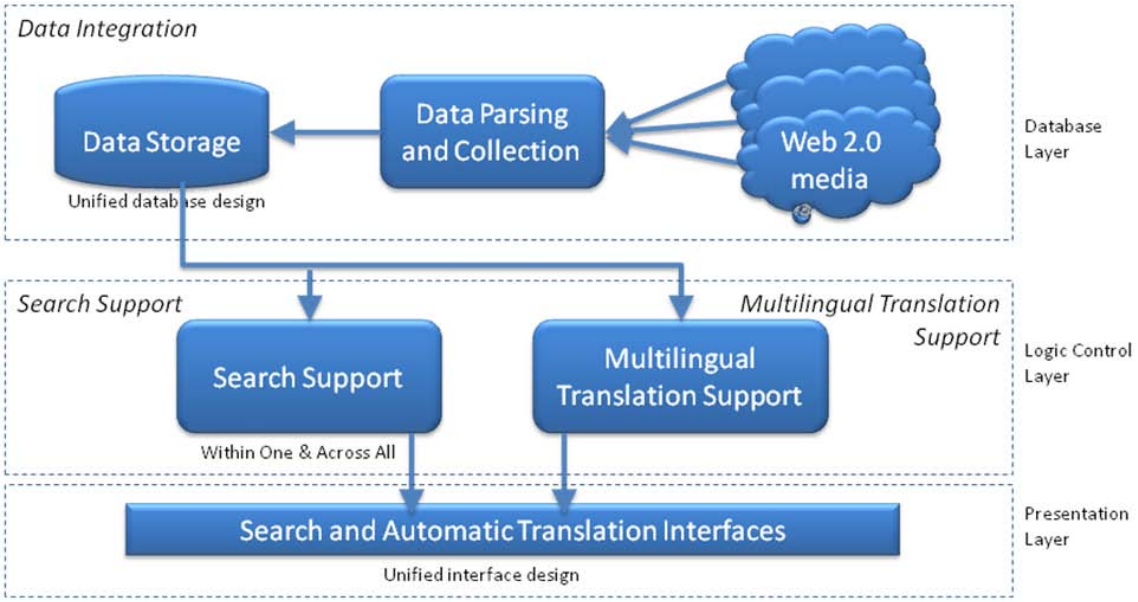
4. Design and Implementation of the Social Media Portal

The portal framework adopts a three-tier architecture. As shown in Figure 1, the architecture consists of a database layer, a logical control layer, and a presentation layer, from bottom to top. To address the data integration and multilingual issue, three major functions are implemented, including data integration support, search support, and multilingual translation support. The first one is implemented in the database layer, while the other two are implemented in the logic control layer.

4.1 The Database Layer

The database layer deals with data collecting and parsing from targeted social media sites. To collect social media data, both complete spidering and incremental spidering are utilized. Complete Spidering is applied to forums the first time they are added to our collection, while incremental spidering is adopted if the forums already exist in the collection. When a new social media site is first added to the data collection, the complete spidering is applied to collect all available postings.

Figure 1. System framework architecture



Incremental spiders are then designed to identify and collect postings generated after the last updating time of the given site. To keep the data sources of the portal up-to-date, incremental spiders can be set to run periodically (e.g., weekly). Each time, only a small portion of data is collected and added to the portal, making the spidering process much more efficient. Different incremental spidering programs are developed and tailored for different social media sites. Using Web forums as an example, since each forum may contain various discussion themes, an incremental spider first needs to collect and compare the URLs of these themes (they can be treated as sub-forums). For each sub-forum, the incremental spider then needs to check the metadata of each discussion thread to get the date (or time) information of the last update of the thread. Threads updated later than the last update date (or time) will be collected. If it's a new thread, then all the postings within the thread will be collected.

Once collected, detailed data fields (such as posting titles, main posting bodies, authors, and posting dates) of the collated social media data will be extracted from the raw HTML Web pages and stored in a local database. A unified database design is adopted for different data sources. Thus, the data integration functionality is implemented in this layer.

4.2 The Logic Control Layer

The logic control layer contains modules for handling search support and multilingual translation support functionalities. This layer acts as middleware that connects the presentation layer and the back-end database layer.

The search functionality allows users to search postings from each data source by searching keyword(s) in various data fields such as message/thread title, message body, author name, and/or post date. Users can specify the logical "AND" or "OR" operation among the keywords to obtain postings containing all or either of the given keywords. In addition to searching information in one particular data source, the system also enables users to search information across all data sources. To provide users with improved consistency and a more positive user experience, a unified search interface design was adopted for different data sources as well as for the across-all data sources search function.

The multilingual translation support functionality is provided in order to process multilingual content written in various languages. It is implemented using Google Translation

API (http://code.google.com/apis/ajaxlanguage/documentation/#Translation), a machine translation based service. Although there are other multilingual translation services available online, we chose the Google Translation API for the following reasons. First, as one of the most widely used online translation services, it provides relatively stable, fast, and accurate translation performance in general. Second, it is free of charge and easy to access. No registration or validation is required. In addition, it can be easily integrated into Java-based Web applications.

4.3 The Presentation Layer

The presentation layer provides user interfaces that enable access to specific functionality. JavaServer Pages (JSP) technology is used to develop the Web pages. When users search non-English forums, the results will be returned in both the original language and in English. Results in both languages are displayed on the same page using a double-column table format, each column for one language and each row for the same posting. To further assist users to search for multilingual content, when they conduct keyword searches in non-English forums, the system allows users to express their search terms in English even when the forum is mainly in another language (e.g., Arabic). In that case, the search will return matches for both the English terms and the Arabic translations of those terms. In the returned search results, keywords in both languages will be highlighted.

5. Evaluation Study and Results

5.1 Dark Web Forum Portal: The Prototype System

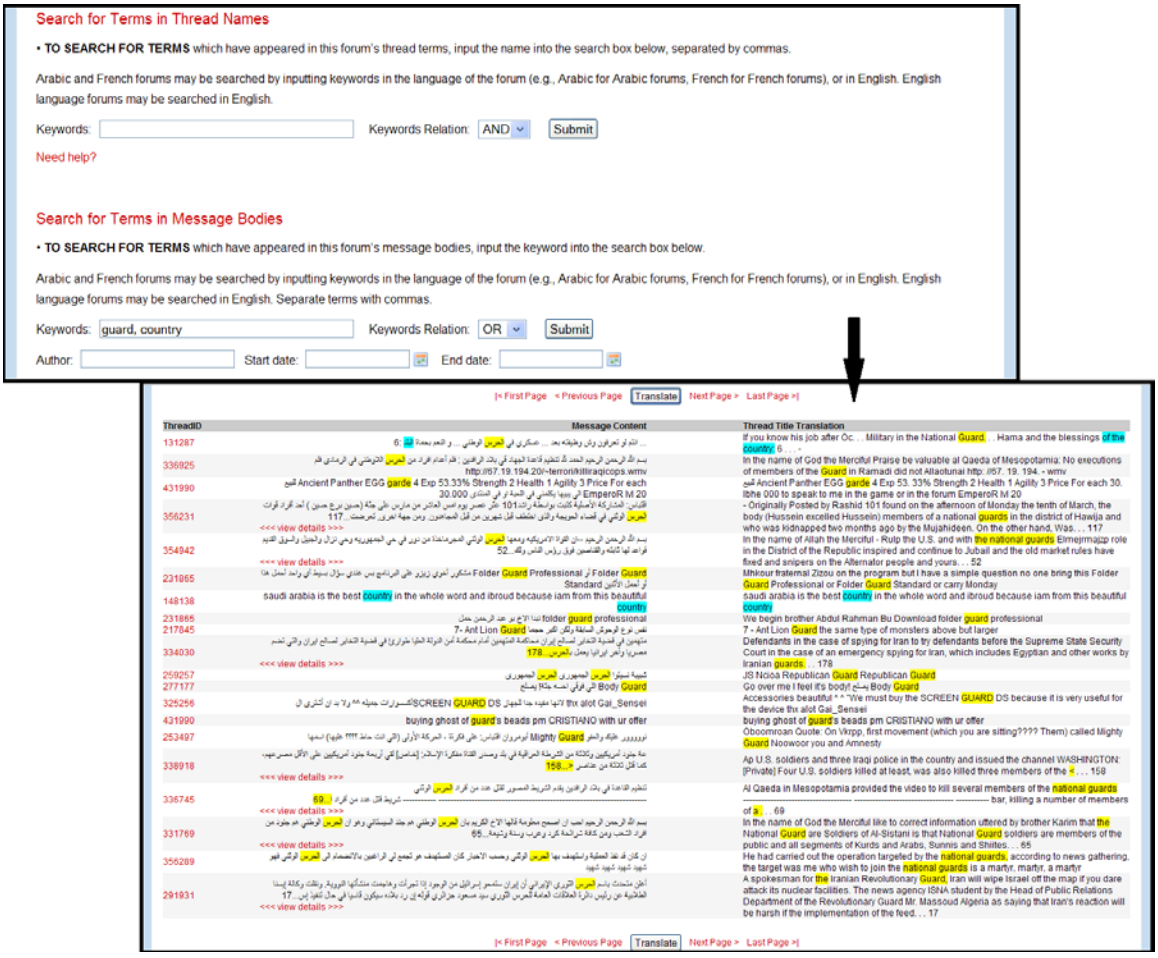
A prototype system, Dark Web Forum Portal (DWFP), has been developed based on the proposed system framework. Currently, the system contains user generated content from 29 important Web forums related to homeland security and selected by domain experts, with a total of 13 million postings from 340 thousand participants. Among them, 17 are in Arabic, 7 in English, 3 in French and 2 in German and Russian. Incremental spidering is set for data collecting and updating. We tested the spidering speed using an Arabic forum as an example – 29,016 new postings were generated by users in a six-week time period, and the spider could collect all of them in 39 minutes (i.e., about 12 postings per second).

The goal of the DWFP is to help users locate and understand and eventually utilize the social media data related to homeland security quickly and easily. It is an infrastructure to integrate heterogeneous forum data, and will serve as a strong complement to the current databases, news reports and other sources available to the research community in this area. Figure 2 shows a screenshot of the search support and multilingual translation support functionalities of the DWFP.

5.2 Benchmark System

A user based evaluation is conducted to assess the performance of the DWFP developed based on the proposed system framework. A big challenge for the evaluation is that there is no existing system providing the same content or functionality as the DWFP. By consulting a couple of domain experts, they mentioned that without an aggregated system as the DWFP, using the search function provided in the original Web forum plus an online stand-alone translation function could be the best way they can leverage to search information in multilingual social media. We therefore used the original Web forum plus the web-based Google translation function (http://translate.google.com/) as the benchmark system. To make a fair comparison, the original Web forum also needs to provide keyword-based search functionality which is similar as the search support provided in the DWFP.

Figure 2. A screenshot of the search support and multilingual translation support functionalities of the DWFP



To minimize potential bias with any single language, two benchmark systems for two different languages were used. One benchmark system is the Arabic Web forum "Alokab," and the other benchmark system is the English Web forum "Islamic Awakening." The Alokab (Islamic Awakening) data collection in the DWFP contains the same set of data as that of the original Web form "Alokab" ("Islamic Awakening"). In addition, both the DWFP and the two original forums provide keyword-based search functions. Since the embedded translation support in the DWFP was implemented using Google Translation API, to make a fair comparison, subjects were asked to use the web-based Google translation function (http://translate.google.com/) when using the original Web forums to complete tasks.

5.3 Hypotheses

A set of hypotheses was developed. By providing consistent data format and user interface of system functionality across multiple data sources as well as integrating the automatic translation capability to the system, we expect that users would be more likely to complete multilingual search tasks successfully and faster when using the DWFP than using the benchmark system. Thus, we hypothesize:

H1: Users achieve higher efficiency when using the DWFP than using the benchmark system.

H2: Users achieve higher effectiveness when using the DWFP than using the benchmark system.

The system capabilities include data integration, search support, and automatic multilingual translation support. In addition, the system also provides some other helpful functions such as allowing users to use keywords in the target language to search information in forums in other languages and the display of messages in both languages with the keywords in both languages highlighted. With these supports, we expected that users would perceive higher system quality, ease of use and usefulness associated with the DWFP compared with the benchmark system. Therefore, we hypothesize:

H3: Users perceive higher system quality of the DWFP than the benchmark system.

H4: Users perceive the DWFP to be easier to use than the benchmark system.

H5: Users perceive the DWFP to be more useful than the benchmark system.

Based on all the above dimensions, users could then perceive higher satisfaction and be more willing to use the DWFP than the benchmark system. Thus, we hypothesize:

H6: Users perceive higher satisfaction of the DWFP than the benchmark system.

H7: Users perceive higher intention to use the DWFP than the benchmark system.

5.4 Experimental Design

A repeated factor design was used. The system is the repeated factor at two levels: the DWFP and the original Web form plus Web-based Google translation function (http://translate.google.com/). Each subject used both the DWFP and a benchmark system to complete a set of tasks. For each subject, the order of the system usage was randomly assigned.

5.5 Measurement Variables

5.5.1 Efficiency and Effectiveness

To test the hypotheses, both objective measures and users' subjective measures were used. Objective measures include efficiency and effectiveness. Efficiency was measured as the amount of time a subject took to complete a task. Effectiveness was measured in terms of task performance accuracy (Chung, Chen, & Nunamaker, 2005; Chung et al., 2004; Marshall et al., 2004; Zhou et al., 2006), which refers to how well a system supports the user to complete a search task correctly. It is calculated as: accuracy = number of correctly answered parts / total number of parts.

5.5.2 A Comprehensive Measure of System Quality – QUIS

Users' subjective measures included system quality, ease of use, usefulness, user satisfaction, and intention to use. System quality measures the specific qualities of the information processing system itself (DeLone & McLean, 1992, 2003; Rai, Lang, & Welker, 2002). As the goal of social media portal is to address the issues like data integration and multilingual issues, system quality is a very important factor to measure social media systems. Ease of use has been treated as an

important component of system quality, and previous research used it as the measure of system quality (Doll & Torkzadeh, 1998; Rai et al., 2002; Seddon & Kiew, 1994). To obtain a more comprehensive assessment of system quality, we used the Questionnaire for User Interaction Satisfaction (QUIS) to measure system quality in this study. The QUIS was developed to measure software usability in a standard, reliable and valid way (Chin, Diehl, & Norman, 1988). It provides measurement items that focus exclusively on assessing users' perceptions of their interactions with a computer system from various aspects (Chin et al., 1988; Harper, Slaughter, & Norman, 1997).

After its initiation, the QUIS has been updated and several versions have been released. Currently, the newest version is the QUIS 7.0 (http://lap.umd.edu/quis/) which contains five major dimensions including: Overall Reactions to the System, Screen factors, Terminology and System Information, Learning Factors, and System Capabilities (Harper et al., 1997). The QUIS 7.0 also contains several optional dimensions, such as technical manuals, on-line tutorials, multimedia and teleconferencing, which can be used to assess computer systems with certain capabilities (Harper et al., 1997). The five core dimensions were used in this study to measure system quality. The optional dimensions were excluded because they are not suitable for the system framework proposed in this study.

5.5.3 Other Subjective Measurement Variables

The measurement items of ease of use, usefulness, user satisfaction, and intention to use were adapted from previously validated scales (Bhattacherjee, 2001; Davis, 1989), with minor wording changes appropriate for the subjects and contexts. All the items used a seven-point Likert scale, with 7 being "strongly agree," 4 being "neutral," and 1 being "strongly disagree."

5.6 Subjects

Subjects were senior level students from a police university in Taiwan. After graduation, most of them either would become police officers or security analysts. They were the potential target users of the DWFP. Subject recruiting and data collection were conducted and provided by collaborator researchers from Taiwan.

5.7 Tasks

Four scenario-based tasks were designed with the assistance of several domain experts, two tasks for searching information from the "Alokab" data collection and the other two for searching information from the "Islamic Awakening" data collection. Each subject used one system (either the DWFP or the benchmark system) to complete the four tasks and then completed a questionnaire to provide their assessment of the system based on the measurement variables. They repeated these tasks using the other system and completed the questionnaire to assess the second system. The order of the tasks to be performed by each subject was randomized, thus mitigating the potential influences of task sequencing.

5.8 Data Analysis and Results

A total of 78 subjects participated in the study, 67 males and 11 females. The mother language of all subjects was Chinese. Both Arabic and English were foreign languages to them. To ensure the success of data collection, the experiment tasks were provided to the subjects in Chinese. The measurement items in the questionnaire were also translated to Chinese. Back translation was conducted to make sure the quality of the translation. Table 1 lists the aggregated subject information. On average, the subjects were relatively experienced computer users with around 10 years of experience. About half of them were trained to be police officers and the other half were

trained to be security analysts. They had limited English reading and writing ability, and none of them could either read or write in Arabic.

Table 1. Aggregated subject information

Atrribute	**Subjects' Characteristics**
Gender	Male: 67; Female: 11
Age	Mean: 25.01; Std dev: 6.67
Years of using computer	Mean: 9.86; Std dev: 3.16
Training focus	Being police officers: 41; Being security analysts: 37
General computer skill[1]	Mean: 4.08; Std dev: 1.11
English ability[2]	Mean: 3.85; Std dev: 1.25
Arabic ability[2]	Mean: 1; Std dev: 0

Note. [1]Rating scale ranges from 1 to 7, with 1 being very unskilled and 7 being very skilled. [2]Rating scale ranges from 1 to 7, with 1 being "can neither read nor write" and 7 being "can read and write very fluently."

Table 2 summarizes the mean and standard deviation values of all measurement dimensions. On average, a user needed 2.33 minutes to complete a multilingual search task when using the DWFP but spent 3.19 minutes when supported by the benchmark system. On average, a user achieved 61.86% accuracy when using the DWFP to complete multilingual search tasks but had 43.40% accuracy when using the benchmark system. For the subjective measurement dimensions, the average ratings toward the DWFP were all above the midpoint (i.e., score of 4) of the 7-point Likert scale, ranging from 4.48 to 4.86. The average ratings toward the benchmark system were all slightly below the midpoint, ranging from 3.20 to 3.83.

Table 2. Descriptive statistics

Measure	**DWFP**		**Benchmark System**	
	Mean	Std dev	Mean	Std dev
Efficiency	2.33	1.06	3.19	1.67
Effectiveness	61.86%	28.34%	43.30%	25.77%
System Quality - A. Overall Reactions to the System	4.53	1.05	3.34	1.22
System Quality - B. Screen	4.72	1.10	3.58	1.28
System Quality - C. Terminology and System Information	4.56	0.90	3.82	1.05
System Quality - D. Learning	4.48	1.14	3.51	1.28
System Quality - E. System Capabilities	4.79	0.74	3.83	1.09
Ease of Use	4.79	1.20	3.37	1.42
Usefulness	4.86	1.13	3.60	1.45
Satisfaction	4.60	1.14	3.31	1.45
Intention to Use	4.73	1.29	3.20	1.33

Note. Efficiency was measured in minutes. Effectiveness was measured by accuracy. Rating scale for the other measures ranges from 1 to 7, with 7 being the best.

The hypothesis testing results are shown in Table 5.3. All hypotheses were significant (p-values < 0.0001). Users achieved significantly higher efficiency and accuracy when using the DWFP than using the benchmark system. Users perceived significantly higher information quality and system quality (in all five dimensions) of the DWFP than the benchmark system.

Users perceived the DWFP to be significantly easier to use and more useful than the benchmark system. Users perceived significantly higher satisfaction of and intention to use the DWFP than the benchmark system. Users perceived significantly higher individual benefit and social benefit of the DWFP than the benchmark system.

Table 3. Hypothesis testing results

Hypotheses	Measure	p-value	Result
H1	Efficiency	<0.0001	Supported
H2	Effectiveness	<0.0001	Supported
H3a	System Quality - A. Overall Reactions to the System	<0.0001	Supported
H3b	System Quality - B. Screen	<0.0001	Supported
H3c	System Quality - C. Terminology and System Information	<0.0001	Supported
H3d	System Quality - D. Learning	<0.0001	Supported
H3e	System Quality - E. System Capabilities	<0.0001	Supported
H4	Ease of Use	<0.0001	Supported
H5	Usefulness	<0.0001	Supported
H6	Satisfaction	<0.0001	Supported
H7	Intention to Use	<0.0001	Supported

Note. All supported at α=0.01.

6. Conclusions and Future Research Directions

A large amount of unstructured/semi-structured data sources and multilingual content are major issues associated with user-generated social media data. To address these issues, an integrated framework of social media portal was developed. The framework enables an integrated access to various unstructured/semi-structured social media data sources by embedding three major capabilities: data integration, search support, and automatic multilingual translation support. The system framework is generic and can be applied to different domains. Based on the proposed framework, a prototype system, the DWFP, was built. This prototype system can be useful for security practitioners and researchers.

A user evaluation study was conducted to assess the performance of the proposed system framework. The involved subjects were from the domain area of the prototype system. They are the potential users of the system. Therefore, compared with using general subjects, the results of the evaluation studies could be more focused and convincing. The study systematically compared the DWFP with the benchmark system based on a wide range of measures related to various aspects of IS adoption and success. The test results showed that users achieved significantly higher efficiency and effectiveness, and perceived significantly higher system quality, ease of use, usefulness, satisfaction, and intention to use when using the DWFP. This demonstrated the advancement of the proposed system framework.

This study has some limitations that can be addressed in future research. First, the current system framework focuses on providing efficient and effective access to user-generated social media data in order to address the two issues of unstructured/semi-structured data sources and multilingual content. However, other key characteristics of social media data are also important and need to be incorporated into the system functionality design. For example, users leverage social media sites to communicate with others by sharing information and exchanging opinions. Thus, social network analysis can be used to examine their interactions. In addition to search and translation functions, future research needs to incorporate social network analysis functionality into the system.

Another key characteristic of social media is user-generated multimedia data. People can post pictures online and record and post audio and video files. All these types of files contain rich information about aspects of people and their opinions and behaviors. However, the current system framework focuses only on textual based user-generated content. Future research needs to incorporate other multimedia data sources such as sounds, pictures, and videos.

In addition, future research can also incorporate other analysis functions, such as sentiment analysis, automatic summarization, and user interactive visualization functions, into the system.

Acknowledgements

This work is supported by the NSF Computer and Network Systems (CNS) Program, (CNS-0709338), September 2007 - August 2010 and HDTRA1-09-1-0058, July 2009 - July 2012. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation or DOD.

References

Abusalah, M., Tait, J., & Oakes, M. (2005). Literature Review of Cross Language Information Retrieval. World Academy of Science, Engineering and Technology, 4, 175-177.

Bhattacherjee, A. (2001). Understanding information systems continuance: An expectation-confirmation model. MIS Quarterly, 25(3), 351-370.

Chen, H. (2009). AI, E-government, and Politics 2.0. IEEE Intelligent Systems, 24(5), 64-86.

Chen, H. (2010). Business and Market Intelligence 2.0. IEEE Intelligent Systems, 25(1), 68-71.

Cheong, F.C. (Ed.). (1996). Internet Agents: Spiders, Wanderers, Brokers, and Bots. Indianapolis, IN: New Riders Publishing.

Chin, J.P., Diehl, V.A., & Norman, K.L. (1988). In Development of an instrument measuring user satisfaction of the human-computer interface (pp. 213-218). Paper presented at the Proceedings of SIGCHI '88, New York. ACM/SIGCHI.

Cho, J., & Garcia-Molina, H. (2000). The Evolution of the Web and Implications for an Incremental Crawler, Proceedings of the 26th International Conference on Very Large Databases (pp. 200-209). Cairo, Egypt: Morgan Kaufmann Publishers Inc.

Chung, W., Chen, H., & Nunamaker, J. (2005). A visual framework for knowledge discovery on the Web: An empirical study of business intelligence exploration. Journal of Management Information Systems, 21(4), 57-84.

Chung, W., Zhang, Y., Huang, Z., Wang, Z., Ong, T., & Chen, H. (2004). Internet searching and browsing in a multilingual world: An experiment on the Chinese Business Intelligence Portal (CBizPort). Journal of the American Society for Information Science and Technology (JASIST), 55(9), 818-831.

Crystal, D. (2001). Weaving a web of linguistic diversity, Guardian Weekly, http://www.guardian.co.uk/GWeekly/Story/0,3939,427939,00.html (Retrieved Febrary 18, 2011).

Dang, Y., Zhang, Y., Chen, H., Hu, P.J.-H., Brown, S.A., & Larson, C. (2009). Arizona Literature Mapper: An integrated approach to monitor and analyze global bioterrorism research literature. Journal of the American Society for Information Science and Technology (JASIST), 60(7), 1301-1319.

Dang, Y., Zhang, Y., Hu, P.J.-H., Brown, S.A., & Chen, H. (2011). Knowledge Mapping for Rapidly Evolving Domains: A Design Science Approach. Decision Support Systems, 50(2), 415-427.

Davis, F.D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Quarterly, 13(3), 319-340.

DeLone, W.H., & McLean, E.R. (1992). Information Systems Success: The Quest for the Dependent Variable. Information Systems Research, 3(1), 60-95.

DeLone, W.H., & McLean, E.R. (2003). The DeLone and McLean model of information systems success: A ten-year update. Journal of Management Information Systems, 19(4), 9-30.

Doll, W.J., & Torkzadeh, G. (1998). The measurementof end-user computing satisfaction. MIS Quarterly, 12(June), 259-273.

Fu, T., Abbasi, A., & Chen, H. (2010). A focused crawler for Dark Web forums. Journal of the American Society for Information Science and Technology (JASIST), 61(6), 1213-1231.

Global Reach. (2004). Evolution of online populations, http://global-reach.biz/globstats/evol.html (Retrieved Febrary 18, 2011).

Harper, B., Slaughter, L., & Norman, K. (1997). In Questionnaire administration via the WWW: A validation and reliability study for a user satisfaction questionnaire. Paper presented at the WebNet 97, Association for the Advancement of Computing in Education, Toronto, Canada.

Hughes, B., Joshi, I., & Wareham, J. (2008). Health 2.0 and Medicine 2.0: Tensions and Controversies in the Field. Journal of Medical Internet Research, 10(3), e23.

Kaplan, A.M., & Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of Social Media. Business Horizons, 53, 59-68.

Kawamura, R. (2010). Social Media's Impact on BI Starts with Web Data Services, http://kapowsoftware.com/blog/index.php/social-media-impact-on-bi-starts-with-web-data-services (Retrieved Febrary 18, 2011).

Marshall, B., McDonald, D., Chen, H., & Chung, W. (2004). EBizPort: Collecting and analyzing business intelligence information. Journal of the American Society for Information Science and Technology (JASIST), 55(10), 873-891.

O'Reilly, T. (2005). What is Web 2.0? Design patterns and business models for the next generation of software. http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html.

Palvia, S.C.J., & Sharma, S.S. (2007). E-Government and E-Governance: Definitions/Domain Framework and Status around the World, http://www.iceg.net/2007/books/1/1_369.pdf (Retrieved Febrary 18, 2011).

Rai, A., Lang, S.S., & Welker, R.B. (2002). Assessing the Validity of IS Success Models: An Empirical Test and Theoretical Analysis. Information Systems Research, 13(1), 50-69.

Roberts, J. (2011). We Have the Data - Now What??!! A Few Examples of Social Media Analytics, http://www.collectiveintellect.com/blog/we-have-the-data-now-what-a-few-examples-of-social-media-analytics (Retrieved Febrary 18, 2011).

Seddon, P.B., & Kiew, M.-Y.A. (1994). In J.I. DeGross, S.L. Huff & M.C. Munro (Eds.), A partial test and development of the DeLone and McLean model of IS success (pp. 99-110). Paper presented at the Proceedings of the International Conference on Information Systems, Atlanta, GA. Association for Information Systems.

Zhou, Y., Huang, F., & Chen, H. (2008). Combining probability models and web mining models: a framework for proper name transliteration. Information Technology and Management, 9(2), 91-103.

Zhou, Y., Qin, J., & Chen, H. (2006). CMedPort: An integrated approach to facilitating Chinese medical information seeking. Decision Support Systems, 42(3), 1431-1448.

Zhou, Y., Qin, J., Chen, H., & Nunamaker, J.F. (2005). In Multilingual Web Retrieval: An Experiment on a Multilingual Business Intelligence Portal. Paper presented at the Proceedings of the 38th Annual Hawaii International Conference on System Sciences (HICSS'2005).