Decision Support Systems 61 (2014) 126–135

Contents lists available at ScienceDirect

Decision Support Systems

journal homepage: www.elsevier.com/locate/dss

An integrated framework for analyzing multilingual content in Web 2.0 social media

Yan Dang [a,*], Yulei Zhang [a], Paul Jen-Hwa Hu [b], Susan A. Brown [c], Yungchang Ku [d], Jau-Hwang Wang [d], Hsinchun Chen [c]

[a] Computer Information Systems, The W. A. Franke College of Business, Northern Arizona University, Flagstaff, AZ 86011, United States
[b] Department of Operations and Information Systems, David Eccles School of Business, University of Utah, Salt Lake City, UT 84112, United States
[c] Department of Management Information Systems, Eller College of Management, University of Arizona, Tucson, AZ 85721, United States
[d] Computer Center, Central Police University, Taiwan

ARTICLE INFO

Article history:
Received 29 December 2012
Received in revised form 29 December 2013
Accepted 19 February 2014
Available online 26 February 2014

Keywords:
Social media
Web portal
User evaluation

ABSTRACT

The growth of Web 2.0 has produced enormous amounts of user-generated content that contains important information about individuals' attitudes, perceptions, and opinions toward products, social events, and political issues. The volume of such content is increasing exponentially, making its search, analysis, and use more difficult and thus favoring advanced tools that aid in information search and processing. We propose an integrated framework that offers an infrastructure necessary for accessing, integrating, and analyzing multilingual user-generated content from different social media sites. Building on this framework, we develop the Dark Web Forum Portal (DWFP) that supports the gathering and analyses of social media content concerning security. Our evaluation results show that users supported by DWFP complete tasks better and faster than those using the benchmark forum. Participants consider DWFP to be better in terms of system quality, usefulness, ease of use, satisfaction and intention to use.

1. Introduction

The phenomenal growth of Web 2.0 has produced enormous amounts of user-generated content available in various online forums, blogs, and social media sites [30]. Such content contains information about individuals' attitudes, perceptions, and opinions toward various products, services, social events, and political issues [30]. The opinion-rich content enables new knowledge discovery in different domains such as healthcare, education, politics, and security. For example, in security informatics area, the Web has become an essential communication media for international extremist groups who are increasingly using the Internet to promulgate their agendas. Due to the high anonymity, easy access, and huge audience of social media sites (e.g., Web forums), international extremist groups often use them to promote violence and distribute propaganda materials. Thus, collecting and analyzing related information from social media sites can be of great importance and interest to security analysts. However, the sheer volume of user-generated content is increasing exponentially, making effective search, analysis, and decision making more difficult [14] for people who are interested in this type of rich data source such as computer and information researchers, security informatics researchers, social scientists, government agencies, and the general public. To address these challenges, advanced, automated tools would provide an effective response. As Kim et al. [34] pointed out, searching and analyzing social media content is tedious and time-consuming. However, inadequate search of social media content produces incomplete information that in turn can mislead the decision making by individuals or organizations alike [20].

Several challenges remain in developing automated search and analysis support tools. For example, gathering and integrating user-generated content, normally semi-structured or unstructured, across different social media sites, is challenging [22,29,32]. The lack of common formats in user-generated content is also problematic; compared with other online resources, such as news or scientific article repositories, user-generated content often lacks a consistent structural organization. Consequently, when searching for relevant content regarding a particular topic or event, such as posted comments representative of mainstream opinions, people often need to visit multiple social media sites to browse and integrate related content with primitive support, if any. Popular social media sites normally vary in their organization or structure, which further constrains the use of a single access procedure to gather content across sites.

Language introduces another dimension of complexity; the content available in many social media sites is usually created in different languages. According to the Internet World Statistics, more than 70% of Internet users are non-English speaking (http://www.

* Corresponding author. Tel.: +1 928 523 7606; fax: +1 928 523 7331.
E-mail addresses: yan.dang@nau.edu (Y. Dang), yulei.zhang@nau.edu (Y. Zhang), paul.hu@business.utah.edu (P.J.-H. Hu), suebrown@eller.arizona.edu (S.A. Brown), ycku1230@gmail.com (Y. Ku), jwang@mail.cpu.edu.tw (J.-H. Wang), hchen@eller.arizona.edu (H. Chen).

http://dx.doi.org/10.1016/j.dss.2014.02.004
0167-9236/© 2014 Elsevier B.V. All rights reserved.

internetworldstats.com/stats7.htm). As a result, effective search and analysis support must properly address issues surrounding multilingual, user-generated content. Most existing search techniques and analytical tools offer limited utilities for gathering and integrating content created in different languages, partially because they are designed for specific grammatical structures or document organization formats [32]. To support enhanced decision making, automated tools should allow people to collect, store, and analyze voluminous user-generated content in different languages, offer desirable flexibility and nearly real-time access, and provide adequate result presentation designs [12].

To address the abovementioned challenges, this study makes contributions by proposing and developing an integrated, architectural framework that provides an infrastructure necessary for accessing, integrating, and analyzing multilingual user-generated content from various social media sites. The framework adopts a modular, multi-layered architecture and offers advanced functionalities for content integration, search, and multilingual translation. This study also contributes to security informatics research and practice by developing an integrated forum portal system (called Dark Web Forum Portal) based on the proposed system framework. The system offers an effective search and analysis support on volumes of user-generated content obtained from different, highly visible, multilingual Web forums of concern to homeland security. The system could be of great importance and interest to computer and information researchers, security informatics researchers, social scientists, government agencies, and the general public.

We conduct an experiment to evaluate this forum portal system. Our evaluation focuses on user task performance measured by accuracy and time efficiency, system quality, ease of use, usefulness, satisfaction, and intention to use, and includes user-generated content from two high-profile security-oriented forums for comparative purposes. A total of 78 senior students from the National Central Police University, Taiwan voluntarily participated in the experiment. According to our experimental results, participants supported by DWFP can complete tasks better and faster than those using the benchmark forum, consider DWFP better in system quality, usefulness, and ease of use, and exhibit greater satisfaction with and intention to use the system.

Following the design science guidelines [27], we adopt the system build-and-evaluate cycle in this study. We first provide description on how the IT artifact was developed to address the challenges in social media data in Section 2, representing the system "build" part of the cycle. We then evaluate the IT artifact via a quantitative study as presented in Section 3, which is the system "evaluate" part of the cycle. Specifically, in Section 2, we first provide a background overview and highlight several key challenges in searching, integrating, and analyzing multilingual content from multiple social media sites (Section 2.1), and then describe in detail about the development of the proposed framework and the forum portal system for security informatics that aim at addressing the identified challenges (Section 2.2). In addition to system development, a systematic and rigorous evaluation of the system is also crucial to design science research as suggested by the design science guidelines [27]. Therefore, we present the system evaluation in detail in Section 3. We describe the evaluation study in Section 3.1, and discuss the data analysis results in Section 3.2. After that, in Section 4, we discuss the study's contributions and limitations, and point out several future research directions. We conclude the paper with a summary in Section 5.

2. Designing the IT artifact DWFP

2.1. Background overview

The community-oriented, highly interactive socialization and communication capability of Web 2.0 has fostered an exponential growth of user-generated content [37]. While this content contains enormous volumes of individuals' attitudes and opinions about various topics, the content across platforms often differs in structure, format, and language. In this section, we provide an overview of user-generated content and highlight several challenges fundamental to the search and analysis of user-generated content.

2.1.1. Overview about user-generated content

User-generated content means the content created and posted on online social media sites by general Internet users [30]. Web 2.0 enables the creation of a large amount of user-generated content by allowing general Internet users to interact and collaborate with each other on social media sites (such as Web forums, blogs, Twitter, and Facebook). This is a huge difference compared with Web 1.0 when general Internet users could only passively view the content provided by major online publishers and webmasters. Previous research that aims at developing Web portals for different areas has mainly focused on collecting and using Web 1.0 content as data sources for the portals, such as EBizPort [35], Nano Mapper [16], MPEG-7 in Moving Image Collections portal [2], and FedStats Web portal [4]. Web 1.0 resources support information (document) publishing and dissemination, typically controlled and managed by the providers. Thus, the content available in these resources normally follows a particular structure; for example, a digital library of scientific articles uses a consistent format to organize key fields of an article, such as title, journal or conference, authors, abstract, and citations.

While Web 1.0 resources provide "read-only" content in specific structural formats and consistent displays, Web 2.0 resources are mostly generated by a community that creates and consumes the content, often in unstructured formats [28,37]. Social networking, a prominent feature of Web 2.0, enables greater information exchange, opinion sharing, and discussion among individuals on a global scale [37], thus resulting in vast amounts of user-generated content. According to Chen [11], the voluminous, dynamic nature of social media data are valuable because they allow us to tap into the "wisdom of the crowd," thus enhancing decision-making. However, the user-generated content available in social media sites, created in different languages and rapidly disseminated without consistent structure, represents challenges for effective search and analysis.

2.1.2. Challenges with user-generated content

2.1.2.1. Content integration across different social media sites. Most user-generated content lacks structure. According to Roberts [39], approximately 95% of the 1.2 zettabytes of data available in the digital universe are unstructured, with about 70% generated by users in social media. Overall, existing tools offer limited support for accessing the unstructured content available in social media sites and presenting it with displays appropriate for users [22,29,32]. As Kawamura [32] comments, integrating social media data represents a critical challenge in Web 2.0.

Central to integrating user-generated content is providing an adequate structure to user-generated content for easy search and access without destroying its richness, such as an opinion issued about an event, who said what, how different people viewed an incident, and how their views change over time. Terman [43] analyzed the integration of social media data related to business and identified the volume and lack of structure as two critical challenges. The absence of a defined data model that specifies data fields and types in social media data also creates difficulty in populating and storing user-generated content in databases [43].

Efforts have been taken to integrate social media data, typically toward one or several given users or events rather than integrating across all or most users and events. For example, Wang et al. [46] proposed an ontology-based, user-centric approach to keep track of an individual's friends and their activities by integrating social media data from different social networking sites. With their approach, people can merge their friends' accounts in multiple social networking sites to obtain a