Object Architected Design and Efficient Dynamic Adjustment Mechanism of Distributed Web Crawlers

Cheng-Hung Tsai, Institute for Information Industry, Innovative DigiTech-Enabled Applications & Service Institute, Taipei, Taiwan
Tsun Ku, Institute for Information Industry, Innovative DigiTech-Enabled Applications & Service Institute, Taipei, Taiwan
Wu-Fan Chien, Institute for Information Industry, Innovative DigiTech-Enabled Applications & Service Institute, Taipei, Taiwan

ABSTRACT

As the global socialnomics rise, big data makes enterprises face the tremendous tide of data at any time. How to efficiently process, analyze these unstructured data and dig useful information from them has been an issue for every level of enterprises to face and settle. Gartner conducted a survey (Gartner CIO Agenda) over 2300 CIOs worldwide and found out that the business intelligence based on big data has been the primary issue (IBM, 2013). Hence, by understanding the above developing trend of social media, this research is mainly based on the authors’ previously proposed paper in ises face the tremendous ti:Design and Implementation of a Web Crawlers Based in Social Networkst any time. How to efficiently process, analyze these unstructured data and dig useful infed architecture of web crawler. This new architecture is added with the concept of object structure for the design and implementation of the whole system. The authors will also investigate the improved object structure that brings the convenience of system maintenance.

Keywords: Big Data, Data Capture, Distributed Web Crawler, Object Architected, Web Mining

1. INTRODUCTION

In the digital living space of social media, because it is based on information technology, every activity of every user can be record as digital data. Take the statistics published by Facebook as example, in average 0.525 billion users will use Facebook every day and there are 138.7 friends per user. In other hand, as every kind of information technology becomes universal along with the maturation of mobile network, such as digital camera, phones and

DOI: 10.4018/ijitn.2015010105
pads, more and more life activities are record through the social networking sites (Innovative DigiTech-Enabled Applications & Services Institute, Institute for Information Industry [IDEAS, III], 2012). Twitter records people’s fleeting thoughts and talks, Foursquare applied by LBS records places visited by people and Instagram that enables mobile sharing of photos helps people capture the instant photos and videos (Jingtian Jiang et al., 2013). These platforms record internet users’ activities, including the implicit motivations. Therefore how to efficiently collect these data and avoid put too much pressure on these social networking sites is the main issue we focus on.

Previously our proposed centralized distributed web crawler is to use a number of them to collect data at one level of web pages (Cheng-Hung Tsai et al., 2014), which is prone to be blocked by sites, not easy to maintain and drags down the entire performance. As for the data collection, it will easily lose the data. Due to these reasons, this research proposes a design method based on the concept of object structure and make every web crawler cooperate with each other through the concept of distributed depth first collection. This method will greatly cut down the occurrence of site blocking and incomplete data collection, and will also fortify the convenience of system maintenance.

This research is mainly to improve the data collection method of web crawler and the design of job scheduling. By the comparison of experiment results, it shows that the proposed method has a greater data collection rate than before (Precision rate: 99.34%, Recall rate: 97.8%), which sufficiently proves the effectiveness of performance promotion.

The remainder of this paper is organized as follows. Section 2 investigates the related works in the past. Section 3 presents the method and technique of the proposed architecture. The results and analysis of the implementation are drawn in Section 4. Section 5 is our conclusion and future work.

2. RELATED WORK

According to the survey of eMarketer at 2013, the number of global social media users has reached 1.73 billion, which covers the 67.7% of the population of internet users. Among them the monthly active users of Facebook are 1.2 billion, the ones of Twitter are over 0.55 billion, the ones of Weibo and Tencent originated in China are over 0.4 billion (IDEAS, III, 2013). Currently the social media users are mostly young people. There are over 90% of internet users aging from 18 to 29 use social networking sites, but the penetration rate of those over 65 years old is only 40%. However, the penetration rate of senior citizens on social networking sites is increasing rapidly.

According to the statistics of III (2013), domestic enterprises have only 15% structured data, yet the other 85% exists as an unstructured data, such as email, web links, logs/journals of information systems. On the other hand, according to the research of IBM (2013), after the internet becomes universal, the data produced on the internet every two days at present is equal to the sum of data from the year of invention of computer to 2003. Under the rapid and massive increase of digital data, formal information systems/platforms are unable to stand. This causes the big data to become the widely concerned issue.

For this reason, this research proposes a technique of object architected design and implementation of distributed web crawlers and provide the importance to the market with the service model by the proposed technique. In addition, we will collect data by our proposed data and conduct an afterward analysis on these data. Hope that we can help these brand owners by providing them with the instant and entire understanding of behaviour blueprint of every potential consumer, and even generate the forecasting insights. This proposed method not only evolves the related techniques of social data collection, but also includes the concept of persona for the future data analysis, which tries to combine sociology, management sci-
An Expert Panel Approach on Developing a Unified System Authentication Benchmarking Index
[www.igi-global.com/article/an-expert-panel-approach-on-developing-a-unified-system-authentication-benchmarking-index/79279?camid=4v1a](www.igi-global.com/article/an-expert-panel-approach-on-developing-a-unified-system-authentication-benchmarking-index/79279?camid=4v1a)

Nation-Wide ICT Infrastructure Introduction and its Leverage for Overall Development
[www.igi-global.com/chapter/nation-wide-ict-infrastructure-introduction/6461?camid=4v1a](www.igi-global.com/chapter/nation-wide-ict-infrastructure-introduction/6461?camid=4v1a)

Open Source Object Directory Services for Inter-Enterprise Tracking and Tracing Applications
[www.igi-global.com/chapter/open-source-object-directory-services-for-inter-enterprise-tracking-and-tracing-applications/115138?camid=4v1a](www.igi-global.com/chapter/open-source-object-directory-services-for-inter-enterprise-tracking-and-tracing-applications/115138?camid=4v1a)