Research on the Construction of Software for Automatic Data Collection and Processing on the Internet

ICTU Student 176
Thai Nguyen University of Information and Communication Technology
Thai Nguyen University, Thai Nguyen, Vietnam
student176@ictu.edu.vn

***Abstract*—This thesis focuses on the development of software to automatically collect and process data from the Internet. The software aims to gather, categorize, and manage data from various web sources, making it easier for users to access and utilize relevant information quickly. The research covers the methods of web data collection, including search engine operations, PHP programming, and MySQL database management, to ensure the software can run on multiple platforms and provide stable performance. The software was tested by collecting data on weather forecasts, stock markets, gold prices, and exchange rates from various websites. The results demonstrated the feasibility of the approach, but the software was not fully completed.**

***Index Terms*—web data collection, search engine, PHP, MySQL, data processing, web crawler**

I. Introduction

The thesis explores the challenges and solutions in automatically collecting and processing information from the web. It covers the basics of search engines, web crawlers, and the technical aspects of PHP and MySQL, which are essential for building a robust data collection system. The research aims to develop a software tool that can efficiently gather and manage data from multiple web sources, providing users with timely and relevant information. The project focuses on implementing functions to collect data on weather forecasts, stock markets, gold prices, and exchange rates, among other areas [1]. Published books on the methods and subject of this project provide context for this English summary [6], [7], [8].

II. Methods

The methods include a detailed analysis of search engines, web crawlers, and the implementation of PHP scripts for data extraction. The software was tested by collecting data on weather forecasts, stock markets, gold prices, and exchange rates from various websites. The PHP scripts were used to parse and extract data from HTML pages, while MySQL was employed to store and manage the collected data. The software was designed to run on multiple platforms, ensuring its versatility and adaptability. However, the software was not fully completed, and some functionalities were not fully implemented [1].

Supervisor: ICTU Teacher 116, Thai Nguyen University of Information and Communication Technology.

III. Results

The software successfully collected and processed data from multiple websites, including weather forecasts, stock markets, gold prices, and exchange rates. The results demonstrated the feasibility of the approach, but the software was not fully completed. The implemented functions included data collection from weather forecasts, stock markets, gold prices, and exchange rates. The software was tested on various websites, such as the Vietnamese National Meteorological Center, VnExpress, SJC, and Vietcombank. However, due to time constraints and resource limitations, not all functionalities were fully developed, and some features were not fully implemented [1].

IV. Discussion

The research highlights the importance of automatic data collection in managing the vast amount of information available on the web. The developed software can be integrated with other systems and run on different platforms, making it versatile and adaptable. The implemented functions, such as data collection from weather forecasts, stock markets, gold prices, and exchange rates, demonstrate the feasibility of the approach. However, the software was not fully completed, and some functionalities were not fully implemented. Future work could focus on expanding the software's capabilities to cover more areas and improve its efficiency [1].

V. Conclusion

The thesis concludes that the developed software can effectively collect and process data from the web, addressing some of the challenges faced by users in accessing relevant information. Future work could focus on expanding the software's capabilities to cover more areas and improve its efficiency. The software was tested on various websites, such as the Vietnamese National Meteorological Center, VnExpress, SJC, and Vietcombank. However, due to time constraints and resource limitations, not all functionalities were fully developed, and some features were not fully implemented [1].

References

[1] ICTU Student 176, ”Research on the Construction of Software for Automatic Data Collection and Processing on the Internet,” internal undergraduate thesis (original in Vietnamese), Thai Nguyen University of Information and Communication Technology, 2010.

[2] 1 Giáo trình SQL - Trần Nguyên Phong - Đ ại học Khoa học Huế, 2004

[3] 2 Giáo Trình Nhập Môn PHP & MySQL - Xây Dựng Ứng Dụng Web -

[4] Phương Lan - Nhà xuất bản Lao động - Xã hội, 2005

[5] 3 Bài giảng cơ sở dữ liệu – Đặng Thị Thu Hiền – NXB Giao Thông

[6] L. Welling and L. Thomson, PHP and MySQL Web Development, 3rd ed. Sams, 2004. ISBN 9780672326721.

[7] M. E. Davis and J. A. Phillips, Learning PHP and MySQL. O’Reilly Media, 2006. ISBN 9780596101107.

[8] A. Silberschatz, H. F. Korth, and S. Sudarshan, Database System Concepts, 5th ed. McGraw-Hill, 2005. ISBN 9780072958867.