報告番号	甲　第　12800　号

主論文の要旨

論文題目　**A Study on Composite Data Mining Methods for Linking Real-World Information with Web Resources**

（実世界情報と Web 資源を関連付けする統合的データマイニング手法における研究）

氏　　名　廖　宸一

論文内容の要旨

Nowadays, the rapid growth of the Internet of Things (IoT) and the ubiquitous computing make the large-scale real-world data (namely the sensor data) growth. The real-world data are data collected by sensors from the real-world. Sensor data can be self-organizing data via spatial-temporal information. Smartphone as a prominent data collection device can take photos, collect GPS data in build-in camera and GPS sensor. It can produce a photo with shooting location and time. A panoramic camera combined with Light Detection and Ranging (LiDAR) can collect 3D point cloud data. Through calibration, point cloud data with street images can be automatically corresponded.

The Web data are typically represented in HTML documents including Web text and Web images on the Internet. Search engines were invented to organize these mass volume of Web data. Users can search relevant Web images and Web text in keywords. One of the major limitations of current search engines is that users must have enough prior knowledge to provide the appropriate keywords. If sensor data can be used as the seed for search engine, which makes it possible to search Web data even if the user is in an unfamiliar environment. Therefore, the challenge to link sensor data with relevant Web data is a crucial task.

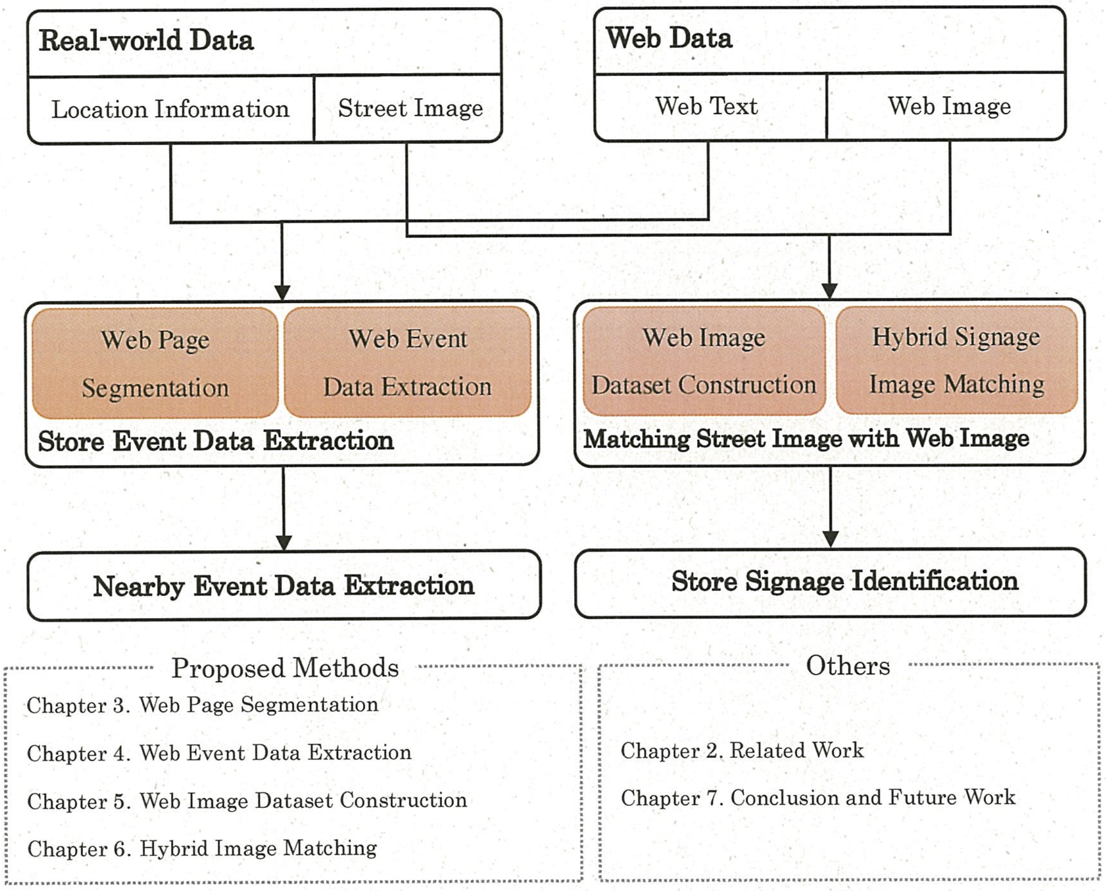
Real-world Data
Location Information
Street Image
Web Data
Web Text
Web Image
Web Page Segmentation
Web Event Data Extraction
Store Event Data Extraction
Web Image Dataset Construction
Hybrid Signage Image Matching
Matching Street Image with Web Image
Nearby Event Data Extraction
Store Signage Identification
Proposed Methods
Chapter 3. Web Page Segmentation
Chapter 4. Web Event Data Extraction
Chapter 5. Web Image Dataset Construction
Chapter 6. Hybrid Image Matching
Others
Chapter 2. Related Work
Chapter 7. Conclusion and Future Work

Figure 1. Outline of This Thesis.

In this thesis, the discussion starts from the point whether it is possible to mine relevant Web data through real-world data. The store information is selected as the research object because the store information typically includes both the real-world data (location, photos, etc.) and Web data (store's Web page etc.). As shown in Figure 1, two types of real-world data are used to link relevant Web data. Firstly, GPS location data is used to extract nearby store event records from Web pages. Secondly, street images taken by users are used to identify store signages and link them with corresponding Web pages.

For the first task, a Web mining scheme for nearby event extraction is proposed in this thesis. As shown in Figure 2, there are two core methods for extracting event data from a Web page. The first proposed method is Web Page Segmentation introduced in Chapter 3 in detailed. This method segments a Web page into records where a record refers to an area of a Web page. Since the page is represented in HTML, this task involves

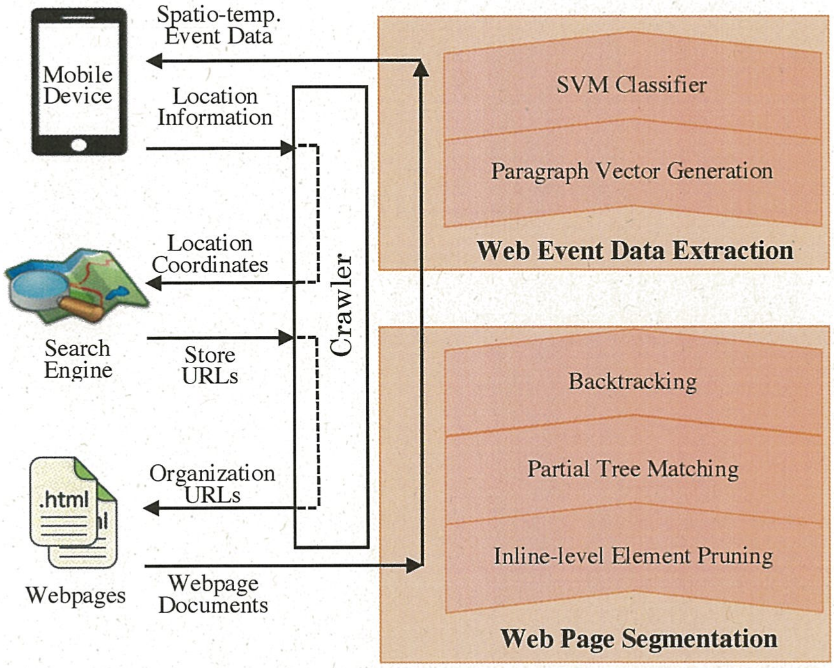

Figure 2. Pipeline of nearby event data extraction from Web pages. Since Web pages of nearby stores were searched and downloaded from search engine according to location coordinates of user device, proposed Web Page Segmentation algorithm converts Web pages into processable structural data and proposed Web Event Data Extraction algorithm extracts event data from them.

HTML code being organized into different records. The second method is event data identification introduced in Chapter 4 in detailed. After Web page segmentation, these records include mass of non-event data. To eliminate the non-event data, the event data items are identified in proposed Web Event Data Extraction method.

For the second task, a store signage identification method based on image-matching is proposed. As shown in Figure 3, this identification process includes two steps. In the first step, a Web Image Dataset Construction method is proposed, which automatically generates target matching database in Web mining technique requiring only a list of store names and their corresponding Web addresses as the input. This method is introduced in Chapter 5 in detailed. In the second step, a Hybrid Image Matching method that combines the deep learning approach with the feature point matching for signage identification is proposed. This method is introduced in Chapter 6 in detailed.

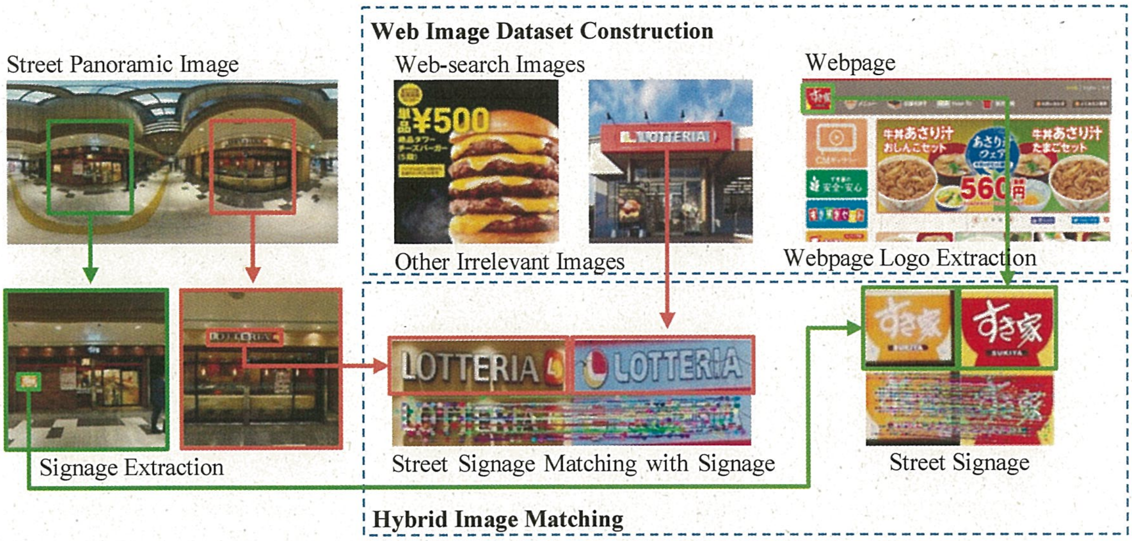

Figure 3. Example of the signage identification process. Two image patches ('LOTTERIA' and 'SUKIYA') are cropped out from street images as query images. The 'LOTTERIA' patch is matched with a similar looking signage patch, which is extracted and cropped from thousands of other Web search images. The 'SUKIYA' image is matched with the official 'SUKIYA' logo, which proposed logo extraction method had effectively extracted from its Web page.

The more detailed introduction of above-mentioned background, motivation, and description of proposed methods is represented in Chapter 1.

In Chapter 2, the related approaches, which aim at the similar tasks, are investigated. These related approaches are analyzed in detailed and compared with the proposed methods. For event data extraction task, the investigations are involved based on Web page segmentation approaches, which generate wrappers to segment a Web page into atomic records, and event data extraction correlation techniques, which discuss how to extract meaning information from a Web page. For signage identification task, because signage is typically matched with text or images in target data resources, the related approaches are organized into groups of textual-matching based methods and image-matching based methods.

In Chapter 3, a Web Page Segmentation method is proposed. A semi-structured HTML document cannot be processed directly. The proposed Web page segment algorithm converts semi-structured HTML document into structured data including segments and records, which correspond to data tables and rows in a relational database, respectively. It can be seen as a reverse engineering of Web page generation. The main process of Web page segmentation is a Partial Tree Matching, which reverses the processing of HTML document generation. Through matching similar HTML structures in an HTML document, it can extract data records from a HTML document. In addition, two algorithms: Inline-level Element Pruning and Backtracking, which can also extract independent Web records that the partial tree matching cannot, are proposed. As a result of an experiment, the proposed algorithm can segment 98.76% of event records from 96 stores in an actual shopping mall. The source code of the proposed Web page segmentation algorithm is published as an open source code on GitHub.

In Chapter 4, a Web Event Data Extraction method is proposed. Atomic Web records are extracted from above-mentioned Web page segmentation method but including massive amount of non-event data. For event data extraction from Web pages, a feasible text classifier is proposed for this task. It is more suitable for classifying short sentences such as Web event records. A F1-score is used for evaluation of the proposed event records classification. As a result of evaluation experiment, it achieves a higher F1-score of 91.61% of event record classification than others. An application Event.Locky is developed to validate the usability of the proposed Web page segmentation and Web event data extraction from Web pages.

In Chapter 5, an automatic Web Image Dataset Construction method is proposed. The proposed method requires only a list of store names and their corresponding Web addresses as the input. The list is easily obtained from official Websites of an experimental shopping mall. The street images from the same shopping mall are input without labeled signage, which is to be identified. In this stage, the naive crawling process yields a massive amount of unrelated information, such as irrelevant images from an image search engine or non-logo images from a store's Website. Several data cleansing methods are introduced to resolve these problems: A two-step method first filters relevant facade images from Web-search images using VGG16 fine-tuning, and then crops the desired signage patches using YOLOv2. Meanwhile, a structure mining method and a statistics-based method extract Webpage logos from Websites. These two methods can effectively reduce the amount of irrelevant information obtained from data crawling. The proposed method can reduce 83.46% and 98.86% of irrelevant images from 55,783 Web-search images and 18,381 Web page images, respectively.

After Web image dataset construction, in Chapter 6, a Hybrid Image Matching method is proposed that combines the deep learning approach with the feature point matching. In this method, a pruning algorithm based on deep learning approach is proposed to match each signage patch from the query dataset to their candidate matching datasets. As each dataset is only a small subset of the total number of images obtained from the Web mining results, this effectively reduces the processing time required during the matching process. The matching process is based on feature point matching and RANSAC. It can effectively cope with noise, unbalanced data from Web mining generated datasets. As a result, the query signage in street images is output with identified store names and Web addresses. Finally, the proposed signage identification method achieves 91% accuracy in a real-life scenario. The source code of the proposed Hybrid Image Matching algorithm is published as an open source code on GitHub.

In Chapter 7, the methods proposed in this thesis are summarized. The contributions in this thesis and the improvements from related approaches are represented. In addition, the shortcomings that can be foreseen in the course of this thesis or the work not done at this stage are pointed out, which should provide inspiration for future works. Because the proposed methods are not way-size-fits-all and have a certain application range. This is to remind readers of the environment and conditions that should be mentioned when using similar methods.