論文の内容の要旨

論文題目

RESEARCH ON LANGUAGE SPECIFIC CRAWLING AND BUILDING OF THAI WEB ARCHIVE

（言語特定クローリングとタイウェブアーカイブの構築に関する研究）

氏名　ソンブーンワィワット　クンワディー

（本文）

The Web has become a new communication medium with historical, cultural, and social significance. Many different organizations, governments, groups, and individuals are increasingly and continually publishing and distributing their information on the Web. Collectively, this makes the Web highly dynamic: web pages are being created, updated, and deleted rapidly. According to recent statistics, seven million new web pages are being created daily; and the average life span of a web page is approximately 44-75 days. It is clear that if we do not preserve the Web, we will permanently lose some important information.

In recent years, many organizations have started Web archiving projects with the aim to preserve the Web information. A notable large-scale Web archiving project is the Internet Archive, which has begun archiving the global Web from 1996. As of July 2007, the Internet Archive has collected web pages corresponding to about 95 billion URLs (1.75 PB of raw data). Other Web archive projects are led by the national libraries of many different countries such as Australia, Sweden, United Kingdom, and Japan. The national Web archiving project is aiming at long-term preservation and access of Web information related to a specific country.

The borderlessness of the Web poses difficulties for comprehensive archiving of the Web of a country (national web archiving). Comprehensive Web archiving usually relies

on a web crawler to automatically download a large number of web pages quickly. In the construction of a national web archive, where the primary goal is to comprehensively collect as much as possible web pages related to a specific country, two conventional web crawling methods are usually used i.e. (1) a domain name based restriction, and (2) geographical location based restriction crawling .

In domain name based restriction crawling, the crawler will be configured to limit crawl scope to only web pages from some selected domain names. For example, one of possible crawling methods for building a Thai web archive is to restrict the crawl within Country Code Top-Level Domain (ccTLD) for Thailand, i.e. ".th" domain name. In geographical location based restriction crawling, the crawler will be configured to download only web pages from servers whose physical locations are within a specific country. In this case, building of a Thai web archive can be done by limit the crawl to only web pages belonging to servers whose physical locations are inside Thailand.

This dissertation studies the challenges and issues faced in collecting language specific web pages and building of Thai web archive. Due to a wide varieties of languages and highly variable degree of cohesiveness of same-language web pages in the Web, conventional crawling methods for building web archives (i.e. domain name based restriction and geographical location based restriction crawling methods) are not suitable for the construction of a comprehensive, high-precision language specific web archive. A more realistic and suitable solution might be a language specific crawler.

In this work, we propose and design a method for crawling language specific web pages without any constraints on domain names and locations of web servers. A language specific crawler is implemented and uses in the building of our Thai web archive. Then, we conduct comprehensive link-based and content-based analyses on the Thai web archive derived from our Thai web crawls.

In link-based analysis, we analyze the characteristics and various statistical properties of the Thai Web graphs associated with Thai web snapshots in the archive. The study can be divided into three levels of abstractions: page-level, host-level, and community level link analysis. We also discuss and suggest some interesting applications of the statistics of Thai Web graph e.g. comparison of spam diffusion rates between different Web sub-regions (such as Thailand vs. Japan), and the discovery of new crawl seeds by studying the linguistic purity of web communities. In content-based analysis of the Thai web archive, we will focus on the detection of semantically meaningful socio-topical web keywords and their evolution patterns.

Following, we will give a detailed outline of the dissertation.

The dissertation starts by studying how we can efficiently collect web pages written in a specific language from the borderless Web space. Towards this goal, we first identify hyperlink patterns that frequently lead to Thai web pages by analyzing characteristics and graphical structure of a large Thai web snapshot. Based on the guidelines derived from linguistic analysis of the hyperlink patterns of the Thai Web mentioned earlier, we devise page-level and server-level language specific web crawling methods for Thai web crawling. Because web crawling consumes resources of remote servers, it is socially, economically, and ethically inappropriate to test a crawler on the real Web. To overcome this difficulty, we have designed and implemented a trace-driven web crawling simulator which utilize large real crawl log to simulate the real Web on a single local machine.

The evaluation of the proposed language specific crawling methods is done on the crawling simulator. The simulation-based evaluation results show higher performance of our proposed methods than conventional crawling strategies. The crawling method with the highest precision and coverage is the server-level strategy. We then implement our language specific web crawling method on a language specific crawler which will be used in the building of Thai web archive.

Unlike previous works on web archiving whose primary concerns are long-term access and preservation of the Web information, this work focuses on deriving values from the archives. The remaining parts of the dissertation deal with analysis and mining of the Thai web snapshots, and discuss how we can utilize the obtained statistical properties in future crawls, web archive managements, and spam detection. We analyze and mine the Thai web snapshots stored in the Thai web archive using both link-based and content-based techniques.

In link-based analysis and mining, we study several statistical properties of the Thai Web graph such as degree distribution, connectivity, and large-scale structure. The study can be divided into three levels of abstractions: page-level, host-level, and community level link analysis. For each level of abstraction, we try, as much as possible, to compare the derived characteristics of the Thai Web graph with other sub-regions of the Web. We also apply a web community extraction algorithm to the Thai web snapshots. We study these web communities in many aspects such as comparison with a real-world Web directory, linguistic purity of web community, and the evolution of some socially significant Thai web communities.

The statistical properties derived from link-based analysis of Thai Web graph can be used as a feedback for improving crawling strategy, managing of the web archives, and developing novel link-based algorithmic tools. Regarding the utilization of statistical

results obtained from our link-based analysis of the Thai Web, we discuss some interesting applications of our statistics e.g. (1) degree distribution analysis for spam detection, and (2) linguistic purity of web community for crawl seeds expansion.

In content-based analysis and mining, we study the evolution of Thai web keywords and explore its relationship with real-world social events. As the Web is now being inundated by hyperlinked information issued by many organizations around the globe. Current events and trends that are happening in the real world may be detected from the Web. We first study the statistical characteristics of socio-topical web keywords sampled from Thai web archive. The socio-topical web keyword is a keyword relating to some topics of interest in a real-world society. We propose a method for extracting these socio-topical keywords from a series of web snapshots. Our proposed method relies on the correlation between link-based and contend-based characteristics of meaningful topical web keywords. By studying the evolution patterns of the extracted socio-topical keywords, it is possible to detect an event and/or trend which were/are happening in the real world.

Finally, the dissertation ends with a summary of main results and a discussion of the future work and remaining open problems