Thai Related Foreign Language Specific Web Crawling Approach

Tanaphol Suebchua, Bundit Manaskasemsak, and Arnon Rungsawang

Massive Information & Knowledge Engineering
Department of Computer Engineering, Faculty of Engineering
Kasetsart University, Bangkok 10900, Thailand.
job,un,arnon@mikelab.net

Abstract. National web archives have been successfully made available through domain- and language-specific web crawlers for years. We here propose another focused web crawler for collecting foreign language web pages that are also related to a nation. Rather finding the most relevant web pages, an ensemble machine learning has been trained with selective features to find relevant clusters of unvisited web pages, called website segments. During consecutive crawling cycles, the machine will be retrained with features extracted from new found website segments. Preliminary experiments in the real web space on Thai-tourism related topics show that this approach can take advantage of recent crawling experiences to produce more promising harvest rates than traditional breadth- and best-first baselines.

Keywords: web archive, topical crawler, focused crawler, website segment, ensemble machine learning

1 Introduction

National web archives preserve national knowledge and cultural information for generations to come. To build one, we have to gather web pages which are related to a specific nation as many as possible [1–5]. Those target web pages can roughly be categorized into 3 groups, i.e., (1) web pages which belong to a national domain name, (2) web pages which are written in a national language, and (3) web pages whose contents are written in other foreign languages, but related to the nation. For the first two groups, researchers successfully utilized a domain-specific web crawler [3, 5–7], and a specific type of focused crawlers, called the language-specific web crawler [8–11], to gather the web pages. However, the third group of foreign language web pages are uncovered by those two formers. The missing web pages may contain informative data, such as thought, aspect to a country, or useful information for foreigners. For example, English web pages which contain information about Thai tourism attractions would be beneficial for all foreign travelers who interest to visit Thailand.

In this paper, we rather consider to localize a website segment, i.e., the subset of web pages, than an individual one. We hypothesize that each already downloaded (or source) segment can give helpful clues to predict the relevancy of

T. Herawan et al. (eds.), *Proceedings of the First International Conference on Advanced Data and Information Engineering (DaEng-2013)*, Lecture Notes in Electrical Engineering 285, DOI: 10.1007/978-981-4585-18-7_72,

unvisited (or destination) ones. The set of selective features are extracted from the source segments to train an ensemble classifier to prioritize the destination segments in the current crawling frontier. During consecutive crawling cycles, the machine updates its knowledge with the new found segments. Preliminary results on the set of Thai-tourism related topic web pages written in English show that this approach provides better harvest rate than the traditional ones.

2 Related Work

At early day, web engineers simply used domain-specific web crawlers to download web pages within the corresponding country code top-level domain to build their national web archives [1–5]. However, those simple crawlers miss many targets ending with .com, .net, .org, etc. Researchers then used another type of web crawler, called the focused crawler, to collect national related web pages. For example, Somboonviwat et al. [9] proposed a focused crawler, called the language-specific crawler. They first detected the language of web pages, and then developed a set of heuristic rules concluded from the observation of link characteristics in a small sample set of Thai web graph to direct their crawler to the unvisited targets. Srisukha et al. [10] proposed to use the machine learning to build the language-specific crawler, while Tadapak et al. [12] proposed a framework to predict the relevant website rather than an individual web page.

3 Relevant web page identification

In this paper, we define a target web page as a web page whose textual content is related to a Thai-tourism topic, and is written in English. To build a training dataset, we manually select some seeds from the ODP [13] of following categories: (1) Thai tourism, (2) foreign country tourism, and (3) non-tourism. We then launch a breadth-first crawler to collect at most 300 pages per website, and use the LangDetect library [14] to select only English web pages. We extract word-based feature vectors; ones from the first category are marked as positive examples, the remainders are marked as the negative ones. To train a Naïve Bayes classifier for relevant web page identification, we run 10-fold cross validation 10 times, and finally choose the best classifier with 95.78% geometric mean value.

4 Thai-related foreign language specific web crawler

Our crawling architecture, depicted in figure 1, composes of three main components: a Segment Identifier, a Segment Predictor, and a Segment Crawler.

4.1 Segment Identifier

The observation concluded from the training data set reveals that many target websites host only a small cluster of relevant web pages, and their URLs mostly

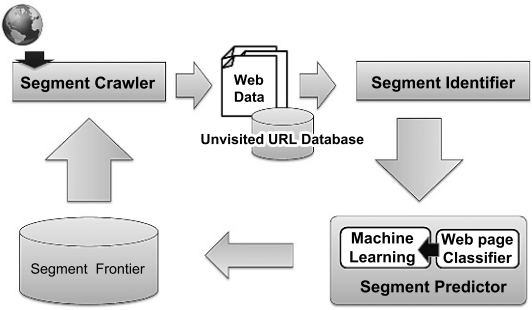

Fig. 1. The architecture of the Thai-related foreign language web crawler.

share the common prefixes. We then design our Segment Identifier to group web pages which share the same longest logical directory path, i.e., the website segments. To give an example of website segment notion, figure 2 illustrates a sample set of URLs from http://www.kpnews.org in which we can later group them into four segments.

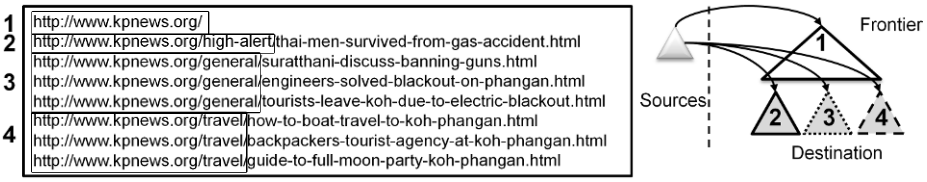

Fig. 2. Sample website segments in http://www.kpnews.org

4.2 Segment Predictor

Segment Predictor is a machine learning based classifier which will be trained to predict the relevancy of the destination website segments. At present, we observe the website segment's characteristics, and extract six features from the source segments in which there is at least one link pointing to the destination segment under consideration as follows.

Geolocation feature: We hypothesize that a destination segment should be relevant if it has been referenced by the source segments whose geolocations are often found in a certain set of countries. We then construct a geolocation feature vector by (1) counting the number of relevant and irrelevant source segments whose geolocations locate in the top-5 and the other countries of which the

source segments are often found to link to the relevant destination segment, (2) the geolocation of the destination segment.

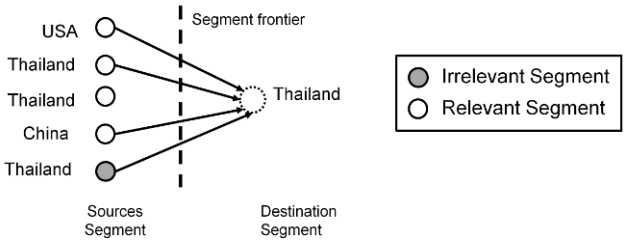

Fig. 3. Example of a geolocation feature extraction.

Suppose that, the top-5 geolocations are USA, Thailand, Japan, France, Russia, respectively, a geolocation vector feature extracted from a sample graph in figure 3 could be written as 1|2|0|0|0|1|0|1|0|0|0|0|TH. This vector means that the destination segment is cited by 1 relevant segment from USA, 2 relevant and 1 irrelevant segments from Thailand, 1 relevant segment from the other country, and the destination segment is located in Thailand, respectively.

Domain name feature: Following the same idea of the geolocation feature, we construct the domain name feature by (1) counting the number of relevant and irrelevant source segments whose the country code top-level domain names locate in the top-5 and the other domains of which the source segments are often found to link to the relevant destination segment, (2) the domain name of the destination segment.

Relevance degree feature: Using the classifier explained in section 3, we can average the percentage of relevant degree of web pages residing within a segment. We here hypothesize that a target segment should be recommended by many high relevant degree source segments. We then construct this feature vector by using the top-5 relevant degree values of the source segments. For example, if the relevant degree of the source website segments in figure 3 are 86%, 78%, 85%, 73.2% and 67% respectively, therefore, the extracted feature vector can be written as 86|85|78|73.2|67.

In-degree feature: We also hypothesize that a target segment would be linked by many relevant source segments. We then construct the in-degree feature by counting the number of relevant and irrelevant source segments.

Anchor text feature: Anchors and their surrounding text in a web page can provide contextual clue about the destination segment. In this work, we propose two methods to construct this feature. The first one is to use word occurrence

extracted from the anchor texts and their 100-characters surroundings. The second method rather extracts N-gram feature from the anchor texts and their surroundings.

6) URL feature: Following the same idea of the anchor text feature, we also construct the URL feature vector by counting either word or N-gram occurrences extracted from the destination segment URL, excluding special characters, "http://" and "www".

4.3 Segment Crawler

Segment Crawler is responsible for downloading web pages within an assigned segment. Website segment with higher relevant score will be first dequeued from the segment frontier database. Then, web pages within that segment will be downloaded. From a downloaded web page, new URLs are extracted and stored in the Unvisited URL Database (cf. figure 1). To avoid downloading too many irrelevant pages from a low relevant segment, we define a discard segment threshold, S. When the Segment Crawler consecutively downloads S irrelevant web pages, it will stop downloading further web pages from that segment.

In order to help the crawler adapt to the new found web segments, we also select all false positive website segment samples and sampling some true positive samples equally at the end of each crawling iteration. Those selected samples will be used to retrain the classifier model of the Segment Predictor later.

5 Experiments

We choose the Thai-tourism related topics as the relevant target web pages. We will first explain how we prepare the training dataset, and then show the crawling result concluded from the Internet setting.

5.1 Training dataset

From around seventy unseen Thai-tourism seed URLs manually selected from the ODP [13], we first use the Google to find their backlinks. We then group those backlinks into website segments, and use the Segment Crawler to download them. From the new list of downloaded URLs, we regroup them into segments and use the Segment Crawler to download web pages in those segments within 2-hops range. Finally, all download web pages are regrouped into segments, and the feature vectors are extracted.

Setting the relevant degree threshold to 50%, i.e., a relevant website segment must compose of at least 50% relevant web pages, we obtain 1,264 relevant and 2,950 irrelevant website segments for the training dataset. We then use the under-sampling method [15] to build the 10-fold cross validation training sets, and explore all simple combiner functions [16] with the ensemble machines. We finally obtain the best predictor model (i.e., 94.8% geometric mean value) using

the Average combiner with the following setting; the link-based features which have been trained with the Naïve Bayes classifier, the word-based anchor text and URL features which have been trained with the Naïve Bayes Multinomial classifiers. Figure 4 depicts the internal architecture of our Segment Predictor.

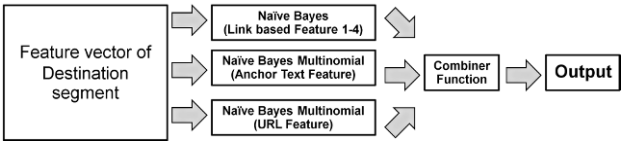

Fig. 4. The internal architecture of a Segment Predictor.

5.2 Internet Evaluation

To evaluate our crawling approach, we compare the following crawlers in the real web space.

- Breadth-first crawler.
- Best-first crawler [17] which follows the destination URL whose parent web page is the most related to a Thai-tourism topic first.
- Our Thai-related Foreign Language specific web Crawler (TFLC-L) which employs only link-based classifier and retrains with the new found segments during each crawling cycle.
- Thai-related Foreign Language specific web Crawler with classifier ensemble-based Segment Predictor (TFLC-ENS) which retrains with the new found segments during each crawling cycle.
- Simple Thai-related Foreign Language specific web Crawler (TFLC-S) which also employs an ensemble-based Segment Predictor but it will not retrain with the new found segments.

We first manually select eight relevant Thai-tourism URLs from the Google, and launch the crawlers from those seed set. We set the discard segment threshold, mentioned in section 4.3, $S = 2$. All crawlers have been restricted to download at most 300 web pages per website.

Figure 5 shows the harvest rate graph concluded from each crawler within one hundred thousand downloaded pages. It can be seen that our proposed approaches provide much better performance than breadth- and best-first crawlers. In other words, our approaches better focus themselves on the relevant web pages region than the baselines.

It can also be seen that there are many ripples in the harvest rate graph produced by our crawling approach. This is because our crawlers always visit the website segments which has the high probability to host relevant web pages

first. Thus, at the beginning of the crawling cycle, many relevant web pages found from those website segments cause the harvest rate to increase. After crawling for a while, crawler will find more irrelevant web pages from the lower probability website segments. This will cause the harvest rate to decline and cause a ripple down in the graph.

When comparing our proposed methods, the crawling performance of TFLC-ENS is slightly better than TFLC-L and TFLC-S. This is because our classifier ensemble-based Segment Predictor can predict the relevant website segments more accurately than using only a link-based classifier. Furthermore, it can also be seen that the harvest rate of TFLC-ENS improves gradually after several crawling iterations too. This shows that the TFLC-ENS can better learn from the crawling experiences than the others. Therefore, the TFLC-ENS would be preferable to use for the large-scale crawling. For the small-scale crawling, the TFLC-S may be much preferable since it consumes less resource than TFLC-L and TFLC-ENS during their classifier update in each crawling cycle.

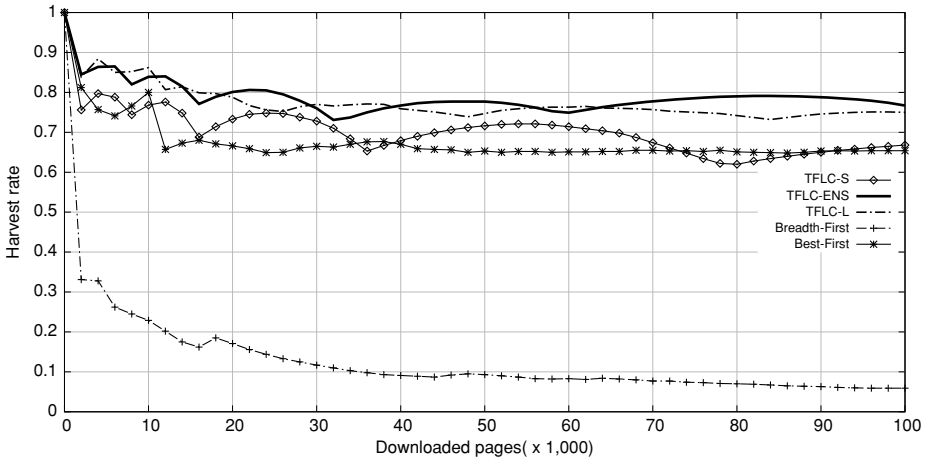

Fig. 5. Harvest rate result.

6 Conclusions

In this paper, we propose another crawling approach for collecting Thai-related web pages which are written in English. We extract several features from downloaded sources website segment to train an classifier ensemble-based Segment Predictor to predict whether the destination website segment could host relevant web pages. Web pages within the website segment with the highest probability values will be later downloaded by the Segment Crawler. Furthermore, in order

to help the crawler adapt to the new environment, new found website segments will be used to retrain the Segment Predictor at the end of each crawling cycle. According to the experimental result on a Thai-tourism web dataset extracted from the real web space, this proposed crawling strategy provides better harvest rate than the breadth-first and best-first baselines. For the future work, We plan to observe the performance on other Thai-related topics, e.g., Thai food, Thai education, etc. We anticipate to exploring the effect of our Segment Crawler parameters to the harvest rate of the crawler too. For the Segment Predictor, we also plan to find more pertinent features and test them with more advance combiner functions in order to archive better harvest rate.

Acknowledgment

The first author thanks the JSTP-NSTDA Thailand for the funding support.

References

1. British Library: UK web archive. http://www.webarchive.org.uk (2011)
2. National Diet Library: Web archiving project. http://warp.ndl.go.jp (2011)
3. Baeza-Yates, R., Castillo, C., López, V.: Characteristics of the web of spain. Cybermetrics **9**(1) (2005)
4. Christensen, N.H.: Preserving the bits of the danish internet. In: Proc. of the 5th IWAW. (2005)
5. Gomes, D., Nogueira, A., Miranda, J., Costa, M.: Introducing the portuguese web archive initiative. In: Proc. of the 8th IWAW. (2008)
6. Baeza-Yates, R., Castillo, C., Marin, M., Rodriguez, A.: Crawling a country: Better strategies than breadth-first for web page ordering. In: Proc. of the 14th WWW. (2005)
7. Bordino, I., Boldi, P., Donato, D., Santini, M., Vigna, S.: Temporal evolution of the uk web. In: Proc. of the 8th ICDMW. (2008)
8. Alabbad, S.H., Alanazi, S.: Language based crawling: Crawling the arabic content of the web. In: Proc. of the ICOMP'09. (2009)
9. Somboonviwat, K., Tamura, T., Kitsuregawa, M.: Finding thai web pages in foreign web spaces. In: Proc. of the 22nd ICDEW. (2006)
10. Srisukha, E., Jinarat, S., Haruechaiyasak, C., Rungsawang, A.: Naive bayes based language-specific web crawling. In: Proc. of 5th ECTI-CON. (2008)
11. Tamura, T., Somboonviwat, K., Kitsuregawa, M.: A method for language-specific web crawling and its evaluation. Systems and Computers in Japan **38** (2007)
12. Tadapak, P., Suebchua, T., Rungsawang, A.: A machine learning based language specific web site crawler. In: Proc. of the 13th NBiS. (2010)
13. DMOZ: Open directory project (ODP). http://www.dmoz.org (2011)
14. Nakatani, S.: Language detection library for java. http://code.google.com/p/language-detection/ (2010)
15. Garcia, S., Herrera, F.: Evolutionary undersampling for classification with imbalanced datasets: proposals and taxonomy. Evolutionary Computation **17-3** (2009)
16. Ranawana, R., Palade, V.: Multi-classifier systems: Review and a roadmap for developers. International Journal of Hybrid Intelligent Systems **3** (2006)
17. Menczer, F., Pant, G., Srinivasan, P.: Topical web crawlers: Evaluating adaptive algorithms. ACM Transactions on Internet Technology **4**(4) (2004)