Enhancing Focused Crawler through Genre Detection

Jiayi QIAN† Tanaphol SUEBCHUA‡ Hayato YAMANA‡

†Waseda University, 3-4-1 Ookubo, Shinjuku-ku, Tokyo, 169-8555 Japan

E-mail : †ula666@fuji.waseda.jp, ‡{tanaphol.su, yamana}@yama.info.waseda.ac.jp

Abstract Web crawlers that attempt to download pages related to a pre-defined topic are called focused crawler or topical crawlers. A general problem in focused crawler is to predict the relevancy of web pages with a given topic before downloading. To challenge the problem, in this research, we hypothesize that genre of a web page, such as news and blog, is a useful feature to estimate the relevancy of the web page. Thus, we propose to utilize the genre of the source web pages to prioritize the unvisited web pages in the priority queue. According to our experiment on finding web pages related to Baseball and Gaming topic, the genre feature could enhance the crawling efficiency of the traditional Best-First focused crawler 10%, at most.

Keyword Focused Crawler, Genre Detection, Web Crawling

1. Introduction

Focused Crawler is a web crawler that attempts to download only web pages that are relevant to a pre-defined topic. The main problem in focus crawling is to predict the similarity of the contents of a web page to a pre-defined topic before downloading. This prediction process enables focused crawler to selectively gather relevant web pages as much as possible while downloading irrelevant web pages as less as possible. It consequently makes a focused crawler to utilize resources, e.g., bandwidth and time, more efficiently than general web crawlers such as Breadth-First search [4].

To predict whether an unvisited web page is relevant or not, recent researches use various mechanisms, such as heuristic rule or machine learning, to analyze the content of the downloaded web pages. The analysis prioritizes the unvisited URLs that are out-links of the downloaded web page. Usually, the URL extracted from high relevant web pages will have high priority to be downloaded. However, the prioritization method based on the relevancy of web pages has its own limitation. Suppose that our given topic is chemistry and our crawler visits an entry web page of the chemistry department in collage. In this case, the entry page is classified as high relevant web page to the given topic. However, most of links on this web pages point to many irrelevant web pages, such as web pages of other departments and collage profiles. In this case, it can be seen that relevancy feature is not enough for prioritize the relevancy.

To solve the problem, in this paper, we have an idea to adopt genre of web pages, i.e., the functional usage of web pages, to predict their relevancy. For example, web pages from blog and news websites can be considered as different genres. We here hypothesize that the probability of a web page being relevant depends on the genre of its source web pages. Suppose that 1) our target topic is informative topic, e.g., education or medical, 2) the crawler downloaded two web pages, and 3) the first downloaded web page belongs to image gallery website, whereas another one belongs to blogs website. In this case, the relevant probability of the links extracted from the image-genre web page should be lower than blog-genre web pages since the image-genre web pages are not likely to cite informative web pages. Based on this idea, we propose a new focused crawling method that assigns the priority score of each unvisited destination web page in accordance with the genre of the downloaded source web page.

The outline of this paper is as followed: In section 2, we introduce the related work of our research. Our proposed method is explained in section 3. Experimental process and results are shown in section 4, and we conclude our research in section 5.

2. Related work

2.1. Focused crawler

Chakrabarti et al. [4] first introduced a topic-specific focused crawler which specified a topic by documents instead of using keywords. In their approach, they first use a classifier to predict the relevancy of the downloaded web pages. Then, they utilize a module named distiller. This

module utilizes the estimated relevancy and linkage between downloaded web pages and unvisited ones to calculate the priority score. The unvisited web page that is likely to be an access point to many relevant web pages will be given more priority than others.

In Rungsawang et al. [2], they defined website segment as a group of web pages whose longest directory paths are the same. For example, http://a.com/gaming/dragonq.html and http://a.com/gaming/zeldom.html belong to the same website segment. With the definition of website segment, the authors proposed to train a machine learning classifier for finding relevant website segments instead of relevant web pages. The web pages belonging to the most probable relevant web pages will be downloaded at once in each crawling attempt. Through testing in real web space, their proposed method shows better performance than both Breadth-First crawler and Best-First focused crawler.

2.2. Genre classification

Fiol-Roig et al. [1] proposed an approach for classifying genre of web pages into the following four types:

1) News: a web page providing news (online representation of traditional newspapers) that are close to real time.
2) Blog: a regularly updated web page (a personal online journal with reflections) which is ordered chronologically.
3) Image: a photo gallery on a website.
4) Video: a web page providing videos or a venue for sharing videos among every person on the Internet.

In their classification approach, they extracted 9 numeric features from the HTML code of web page. The extracted features are as follows:

1) Page's text length: number of text characters in the web page.
2) Internal links: number of links belong to the domain of source web page.
3) External links: number of links belong to any domain other than the domain of source web page.
4) Words related to Video: number of appearances of words "video" and "動画" in the text of web page.
5) Multimedia Objects: number of elements representing videos and flash player tag in the web page.
6) Words related to Image: number of appearances of the words "image", "写真" and "画像" in the text of web page.
7) Image tag: number of elements representing image tag in the web page.
8) Words related to Blog: number of appearances of the words "blog", "ブログ" and "日記" in the text of web page.
9) Words related to News: number of appearances of the words "news", "ニュース" and "新聞" in the text of web page.

These features are used to train several classification algorithms. According to the experimental results, Decision Tree, Logistic Tree and J48 Tree models achieve better performance than others.

Simaki et al. [5] proposed their study to identify genre from different text types. They classified texts into four types: short stories, news articles, recipes and dictionaries. A genre clue indicator is built to indicate characteristics of genre within the text, and the results show high accuracy of detecting different genre clue from each other.

In Berger et al. [6], the authors aim to detect genre of web pages in four classes: blog, forum, portals and original websites. They trained the classifier with a combination of textual content, structural three groups of features: n-grams, structural properties and quantitative properties with a set of 80 features and SVM parameters. Finally, their experiment achieves a high precision of 87.5%.

De Assis et al. [7] aims to distinguish syllabus web pages from content web pages. In this paper, author defines web pages containing syllabus as "Other" genre and could be selected by the crawler.

Simaki et al. [5] proposed their study to identify genre from different text types. They classified text into four types: Short stories, News articles, Recipes and Dictionaries. A genre clue indicator is built to indicate characteristics of genre within text, and the results show high accuracy of detecting different genre clue from each other.

3. Proposed method

3.1. Overview

In this paper, we hypothesize that, the probability of a web page being relevant web page depends on the genre of its source web pages. Under this assumption, we propose a focused crawler that assigns the priority score of each unvisited web page based on both relevancy and genre of

the downloaded source web page. The structure of our proposed focused crawler is shown in Figure 1. Two classifiers are adopted in this crawler; one is a relevant web page classifier, another is a web page genre classifier. The relevancy classifier is the same classifier that was proposed in Rungsawang et al.'s work [2]. As for the genre classifier, we adopted Fiol-Roig et al.'s one [1]. For each downloaded web page, we first utilize both classifiers to predict the relevancy and genre of the web page. The results obtained from both classifiers are as follows:

I: The predicted relevancy class. Because the web page classifier is trained to classify whether the downloaded web page is relevant web page or not. Thus, I ∈{Relevant, Irrelevant}.

R: The probability indicating the relevancy of an input web page with a given topic. This result is obtained from relevant web page classifier.

G: The predicted class of genre obtained from genre classifier. In this research, G∈{News, Blog, Image, Video, Other}.

C: The probability indicating a downloaded web page belongs to its predicted genre class in G. This output is also obtained from genre classifier.

We use all prediction results to calculate the priority score of the URLs extracted from the downloaded web page later. The details on our priority scoring function is described in the following section 3.2.

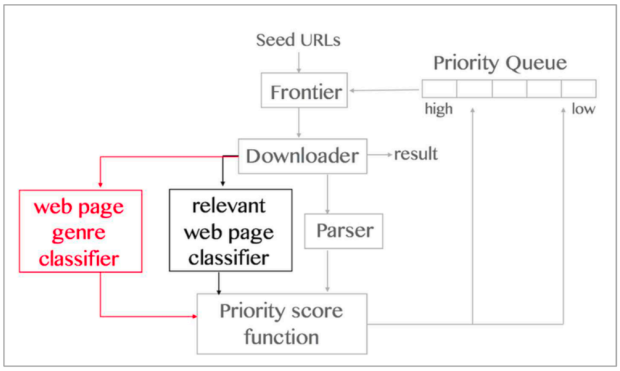

Figure 1: Flow chart of genre based focused crawler

3.2. Scoring function

In this research, we design the priority scoring function as follows:

$$S = w_R \times R + w_G \times (C \times P_{IG})$$

In the above equation, as introduced in Section 3.1, R and C are the probability obtained from relevant web page and genre classifiers, respectively. w_R and w_G represent the weights. The value of w_G is equal to w_G =1-w_R.

P_{IG} shows the probability of a web page being relevant pages depending on its downloaded source web pages' genre G. P_{IG} is calculated by using the following equation.

$$P_{IG} = \frac{N_R}{N_T} \qquad (1)$$

,where N_R is the total number of links to relevant destination pages where the relevancy and genre class of the downloaded source web pages belong to class I and G. As for N_T, it is the total number of source web pages belong to G.

It can be seen that we have to find out every P_{IG} for all possible classes I and G. This can be achieved by analyzing the linkage structure of the subset of the web. The analysis process is later described in Section 4.3.

4. Experiment

According to our investigation on Yahoo! Japan Directory, gaming and baseball topic have highest popularity on the Internet because they have the most number of URLs in the directory. Hence, in this work, we select gaming and baseball topics to evaluate our crawler.

Section 4.1 introduces the web page genre classifier training followed by relevant web page classifier training in section 4.2. In section 4.3, we calculate P_{IG}. The weighting values are calculated in section 4.4. Section 4.5 shows our experimental result.

4.1. Training Web Page Genre Classifier

We classify the genre of web pages based on Fiol-Roig's approach [1]. Besides that four genres defined above, we define "Others" genre that contains any other genre type's web pages.

Among all of the classification methods, we select Random Forests due to its short run-time and ability to deal with unbalanced data. Usually, Random Forest starts with decision tree, a dataset is input at the top of the tree and as it goes down the tree, the data gets into smaller sets that are easily to be analyzed. Random Forests consist of a group of decision tree classifiers on different sub samples of the dataset and it uses average number to enhance the predictive accuracy.

Based on the definition of web page genre, we manually picked up 100 URLs from each category '新聞' (news), 'ブログ' (blog), '写真' (image), '動画' (video), and randomly picked up 100 URLs from other categories

under Yahoo Japan Directory as the training dataset for 'others' group. With these 500 datasets, we trained a Random forest classifier under the test mode with 5-fold cross validation, and the accuracy of this web page genre classifier was 85.4%.

4.2. Training Relevant Web Page Classifier

In this work, we follow Rungsawang et al. [2] to build a web page classifier based on Naïve Bayes Multinomial classifier to calculate the similarity of the input web page and the given topic. In addition, we also use this classifier to estimate the efficiency for performance evaluation.

For relevant webpage classifier, the text appeared in the following HTML Tag of an input web page are extracted:

1) TITLE
2) BODY
3) ANCHOR

We manually examine and select URLs from the following categories of Yahoo! Japan Directory for two topics: 'ゲーム' (gaming) and '野球' (baseball), and launch the Breadth-First search (BFS) crawler to collect at most 300 web pages from each website. All of the web pages have been classified by using language classifier derived from the LangDetect library [3] to filter out non-Japanese web pages. After examined all of the downloaded web pages manually, we got 13,401 relevant web pages grouped in positive samples and 15,303 irrelevant web pages grouped in negative samples for gaming topic, and 30,002 positive and 31,237 negative samples for baseball topic. We use these datasets to build the web page classifier model for each topic by using the 10-fold cross-validation. However, it can be seen our dataset is imbalance. Thus, we apply the under-sampling technique to every training fold generated by the cross-validation process. Finally, we got two web page classifier models for classifying the gaming and baseball with 97.84% and 94.07% respectively.

4.3. Calculating P_{IG}

As mentioned in Section 3.2, we have to find out every P_{IG} for all possible classes I and G. To achieve this task, for each topic, we first select out seed URLs from Yahoo! Japan Category as our seeds. In our seed selection process, the URLs of web page that we previously used to trained classifiers were removed. We then launched the Breadth-First crawler to download and follow the links within in two hops from seed URLs. The number of collected web pages is shown in Table 1. For gaming topic, the total number of linkages between web pages is 2,033,538 links. and for baseball topic, the number of linkage is 2,675,069 links. The number of links is much bigger than web pages due to intricate relationship among links in this dataset, for instance, one destination web page has several source web pages.

We run a Best-first crawler that starts from the seed URLs shown in Table 1 to build a web graph. In this crawler, we use the relevant web pages and genre classifier to decide the relevancy and genre of each web page in the web graph. With the linkage structure in a web graph and the output from both classifiers, we can calculate the P_{IG} for every classes I and G. The values of P_{IG} are summarized in Table 2 and Table 3.

Table 1: Structure of web graph

Dataset	Number of web pages (Baseball)	Number of web pages (Gaming)
Seed URLs	82	156
1st hop	4,868	5,914
2nd hop	37,497	42,724

Table 2: Probability of detecting relevant web pages for "Gaming"

Genre	P_{IG}	
	from relevant source	form irrelevant source
Blog	89.2%	10.9%
Image	70.1%	2.7%
News	71.2%	2.3%
Video	64.7%	0.3%
Others	87.3%	1.1%
ave.	82.9%	1.9%

Table 3: Probability of detecting relevant web pages for "Baseball"

Genre	P_{IG}	
	Relevant source	Irrelevant source
Blog	76.4%	3.3%
Image	36.7%	1.5%
News	52.7%	4.6%
Video	69.5%	1.8%
Others	23.6%	0.4%
Total	57.6%	2.6%

From Table 2 and Table 3, we can see "Blog" genre has the highest probability detecting relevant links for both gaming and baseball topic. "Other" genre has the lowest probability for baseball topic, but for gaming topic, it is the second highest one. Also, in baseball topic, the variation of probability for each genre has a huge range from 23.6% to 76.4%. However, in gaming topic, the probability to detect each genre relevant pages are similar to each other. To find out the reason of high probability in "Other" genre, we examined on the dataset. We found that a crowd of web pages that are mostly the official gaming web site was classified in "Other" genre.

4.4. Finding optimal Weight (w_R and w_G)

As mentioned in Section 3.3, there are two factors in our priority scoring function, i.e., R and $(C \times P_{ig})$. Each factor has its own weight, i.e.,w_R and w_G. To find out the weights that make the crawler achieves the best result, we simulate the crawling process on labeled web dataset (mentioned in section 4.3) with score priority score function $S = w_R \times R + w_G \times (C \times P_{ig})$ in 20 combinations of w_R and w_G from w_R equals to 1 to 0 with interval 0.05, and record the coverage rate of each score function with its weight. Area Under Curve is used to compare the efficiency of each function. When w_R equals to 1, priority score function $S = R$ represents the baseline, which means genre feature does not make any effect in sorting the priority order of web pages and our crawler turns into a Best-first crawler. Figure 2 shows the coverage rate of baseball topic. Here is the calculation function of coverage rate:

$$Coverage\ rate = \frac{\#downloaded\ relevant\ web\ pages}{\#total\ relevant\ web\ pages\ in\ the\ dataset} \quad (2)$$

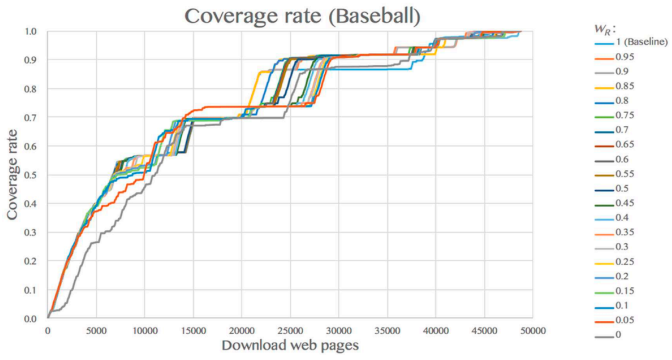

Figure 2: Coverage rate of twenty score function for "Baseball"

Coverage rate represents the ratio of downloaded relevant web pages with total relevant web pages. In baseball dataset, the number of total relevant web pages is 7,124, and for gaming topic, there are 6,901 relevant web pages in total. Due to no discard in the simulating crawling process, each score function achieves 100% of coverage rate in final. However, it is difficult to distinguish which score function shows better performance from this figure macroscopically, we use Area Under Curve(AUC) method to calculate the area each score function reaches. Figure 3 shows the result.

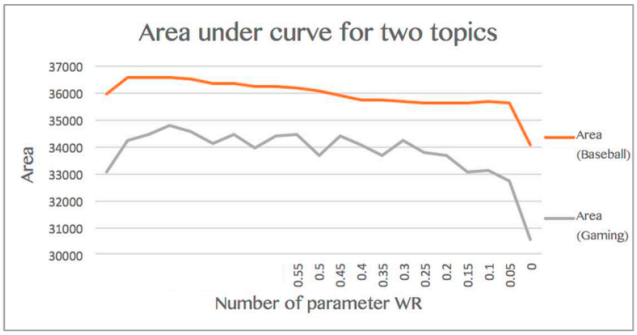

Figure 3: Area under curve for each parameter w_R

For baseball topic, the area declines gradually as parameter w_R decreases from 0.95 to 0.05, when w_R equals to 0.95, the area reaches its maximum of 36,595. On the contrary, the area of gaming topic shows an unstable change as the variation of w_R, the maximum of area is 34,814 when w_R equals to 0.85. For both two topics, most of score functions that added genre feature show better performance than baseline (w_R equals to 1).

4.5. Evaluation on the Internet

For each evaluated topic, we choose the priority function that produces the best result in our crawling simulation and test in the real web space as shown in Table 4.

Table 4: Testing score functions for two topics

	Priority score function (Baseball)	Priority score function (Gaming)
Best-First crawler (baseline)	$S=R$	
Genre enhanced crawler	$S = 0.95 \times R + 0.05 \times (C \times P)$	$S = 0.90 \times R + 0.10 \times (C \times P)$

For each topic, we launched two crawlers with different priority score function under the same environment: initiate crawling at same time with 7 same

seed URLs that are different from our training dataset, which are selected from top google searching results from the keyword "野球"(baseball). The limitation of downloading web pages from same website is set to 300 web pages to prevent duplicated download of same contents. We evaluate the harvest rate of 10,000 web pages downloaded by each crawler. Here is the calculation function of harvest rate:

$$Harvest\ rate = \frac{\#downloaded\ relevant\ web\ pages}{\#total\ downloaded\ web\ pages} \quad (3)$$

Result is shown in Figure 4 and Figure 5. Table 5 concludes the harvest rate of the results of baseball and gaming topic.

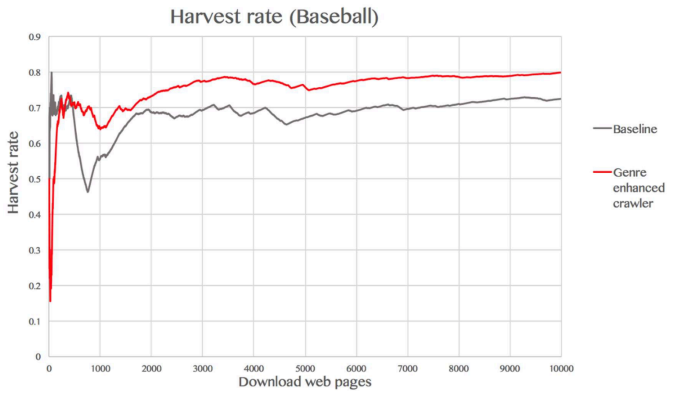

Figure 4: Harvest rate of four different crawlers for baseball topic

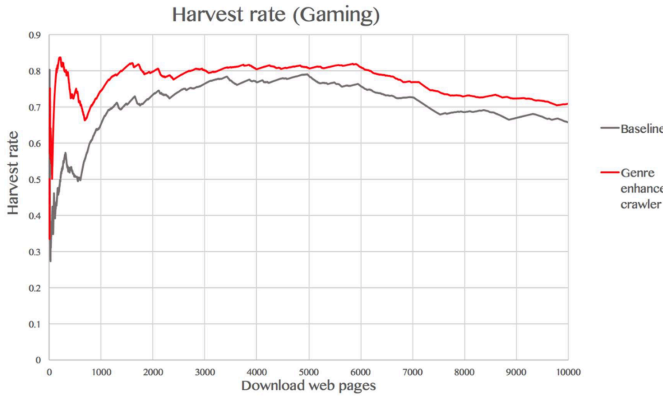

Figure 5: Harvest rate of four different crawlers for gaming topic

Table 5: Harvest rate of Baseball and Gaming topic focused crawler

	Harvest rate (Baseball)	Harvest rate (Gaming)
Best-First crawler (baseline)	71.8%	65.8%
Genre enhanced crawler	78.8%	70.5%

According to the result shown in Table 5, our genre enhanced crawler for baseball topic achieves the highest harvest rate 78.8% after downloading 10,000 web pages. The baseline crawler shows the worst result among these four crawlers. For gaming topic, harvest rates are approximately 70%, which are higher than the baseline crawler achieved 65.8% which utilizes only the relevancy feature.

From the Internet evaluation, we can conclude that focused crawler utilized genre feature shows better performance than normal best-first crawler on baseball and gaming topic.

5. Conclusion

In this paper, we proposed a new topic-specific focused crawling method that incorporates the genre of the web pages, i.e., the type of web pages, to enhance the harvest rate. We assume that the probability, that we will find the links to relevant web pages, also depends on the genre of the source web page. Our method consists of two web page classifier models; The first one is a web page classifier for classifying the relevant web page, whereas the other classifier is used for classifying the five genre of web pages. Our experiments on the Baseball and Gaming topics shows that our genre enhanced Best-First crawler outperforms the simple Best-First focused crawler around 9.7%, relatively.

Although our proposed method achieves better crawling results, there are still some place worth for further improvement. Firstly, in our method, it is not able to adopt the web page genre classifier without pre-analyzing on the topic. Secondly, there are a certain number of relevant web pages exist in web pages belong to "Other" genre. Hence, adding variety of genre is a direction of improvement of our approach. Furthermore, to ensure the generality of our proposed method, we plan to evaluate it on more topics.

References

[1] G. Fiol-Roig, M. Miró-Julià, and E. Herraiz, "Data mining techniques for web page classification," Proc. of 9th International Conference on Practical Applications of Agents and Multiagent Systems, pp. 61-68, 2011.

[2] A. Rungsawang, T. Suebchua, and B. Manaskasemsak, "Thai related foreign language-specific website segment crawler," Proc. of 28th International Conference on Advanced Information Networking and Applications Workshops, pp. 293–29, 2014.

[3] N. Shuyo, “Language detection library for java,” http://code.google.com/p/language-detection, 2010.
[4] S. Chakrabarti, M. van den Berg, and B. Dom, “Focused crawling: A new approach to topic-specific web resource discovery,” Proc. of the 8th International World Wide Web Conference, pp. 1623–1640, 1999.
[5] V. Simaki, S. Stamou, and N. Kirtsis, “Empirical Text Mining for Genre Detection,” Proc. of International Conference on Web Information Systems and Technologies (WEBIST), pp. 733-737, 2012.
[6] P. Berger, P. Hennig, M. Schoenberg, and C. Meinel, “Blog, forum or newspaper? Web genre detection using SVMs,” Proc. of Web Intelligence and Intelligent Agent Technology (WI-IAT), pp. 64–68. 2015.
[7] G.T. De Assis, A.H. Laender, M.A. Gonçalves, and A.S. Da Silva, “A genre-aware approach to focused crawling,” Proc. of the 18th International World Wide Web Conference, pp. 285-319, 2009.