Focused Web Crawlers on Domain-Specific Retrieval Systems

Ika Oktavia Suzanti¹, Fakhrur Razi², Husni³, Eka Mala Sari Rochman⁴, Nurhayati Fitriani⁵

¹,²,³,⁴,⁵ Informatic Engineering Department, Faculty of Engineering, University of Trunojoyo Madura
Email: iosuzanti@trunojoyo.ac.id¹, 160411100058@student.trunojoyo.ac.id², husni@trunojoyo.ac.id³, ekamalasari3@gmail.com⁴, nurhayatifitriani77@gmail.com⁵

Abstract. The need for a large and growing internet has formed a new culture that symbolizes widespread dissemination of knowledge, information and data [1]. Every day hundreds of pages as well as new information were added, deleted and modified. Search engines is a finding information tool. First step of search engines is web crawler, a process crawling webpage to obtain information about its content. Focused web crawler is one of web crawler that scrapping certain relevant webpages to given topic and ignores others that have nothing to do with. Indonesia is an archipelago that number of island reached 17,491, where it makes Indonesia rich of different kinds of travel, culture and food that is characteristic of each region. The amount of information available does not rule out possibility that there are some food recipes do not include staples used so that web crawlers are needed to find out the ingredients. With this application , it is expected to facilitate the search for information on various processed recipes with meat-based ingredients from all regions in Indonesia .

Keywords: Search engine, information tool, focused web crawler.

1. Introduction
The need for a large and growing internet has formed a new culture that symbolizes the widespread dissemination of knowledge, information and data [1]. Every day hundreds of pages as well as new information were added, deleted and modified. Search engine is a finding information tool that consists of two kinds, horizontal and vertical search engine. Horizontal (broad-based) search engines are search engines that provide information for a variety of topics, while vertical search engines only provide information for certain topics [2].

Google, the largest horizontal search engine, in fact has been able to improve user spelling queries, answering majority of questions, even show search results before user presses enter [3]. On the other hand, vertical search engines are also increasingly popular among certain people who are looking for more specific information that cannot be found on horizontal search engines [4]. The first step of a search engine is web crawler, which is a process to crawling webpage to get information about its content. Various types of web crawlers are parallel crawlers [5], distributed crawler [6], topical or focused crawlers [7] incremental crawlers [8], dan hidden web crawlers [9]. Web crawler process begins with embedding where the first url must go then looks for another url based on what was found at that address [10].

Focused web crawler is one of web crawler which scrapping a certain webpages that are relevant to given topic and ignore others that have nothing to do with [11]. Research for focused web crawlers development that have been carried out such as search algorithm optimization [12] [13]the effectivity.
of appropriate keywords [14], and verify web crawler's functionality using model checker [15]. Focused web crawler implementation has been used in various fields including marketing strategies [16], news such as sports news [17] or crime news [18], green building classification [19] and geo-location on social media [20].

Indonesia is an archipelago consisting of 34 provinces with 17,491 islands spread from Sabang to Merauke. Various kinds of tours, culture and food are become a characteristics of each of these regions. The purpose of this research is to create a focused search engine for regional food recipes in Indonesia, especially those that use main ingredients of meat. The amount of information available does not rule out possibility that there are some food recipes that do not include meat as staple used so webcrawlers need to know ingredients used. With this application, it is expected to facilitate search for information on various processed recipes with meat-based ingredients from all regions in Indonesia.

2. Methodology

The system to be developed uses web crawler as a media to retrieve data from the web. In this system, admin input keywords that will be searched then webpage obtained is stored in a database. The data stored in database will be preprocessing text, this process is used for deleting non-letter characters, deleting Indonesian-language stopwords that are built manually for recipe data, and uniforming spelling of terms of each food ingredient used. The data in database is then ranked and displayed when user enters food recipe keywords to be searched.

![Architecture of Meat Crawler](image)

**Figure 1.** Architecture of Meat Crawler

3. Result and Discussion

Website address that is destination of crawlers is a food recipe website including Eresep.com, menuresepmasakan.com, Resepkoki.com, ResepOnline.info and is limited based on the food ingredients used namely meat.

```python
judul=[]
link=[]
deskripsi=[]
urls=[str(i) for i in range (1,2)]
for url in urls :
    page=get("https://www.eresep.com/cari?resep=daging&page="+url)
    soup=BeautifulSoup(page.content, 'html.parser')
    jdl=soup.findAll(class_='entry__title')
    ling=soup.findAll('h2',attrs={'class':'entry__title'})
    judul.append(jdl)
    link.append(ling)
```


After obtaining the appropriate webpage, then webpage is preprocessing text and results are stored in a database.

**Code Program 1. Code program crawling eresep.com**

```python
for a in range (len(jdl)):
    judul+=[jdl[a].find('a').text]
for b in range(len(desk)):
    deskripsi+=[desk[b].find('p').text]
for c in range (len(ling)):
    link+=[ling[c].find('a').get('href')]

Code Program 1. Code program crawling eresep.com
```

### Figure 2. GUI Meat Crawler

Figure 2 is interface of meat crawler application that can be accessed from browser. The system testing is done by calculating precision, recall and F-Measure of several keywords. To a query by one keywords is obtained recall 97%, precision = 98% and F-Measure is 97%, while for queries with two keywords is obtained recall = 82%, precision = 74% and F-Measure = 77%. This shows that application can provide relevant information quite well.

### 4. Conclusions

Based on the results of implementation and testing of meat crawler application it can be concluded that web crawler application can retrieve the intended webpage with a predetermined topic. While based on test results of recall and precision, value of recall of 0.82 and a precision of 0.74 and F-Measure 0.77. The accuracy is relatively small because
the current text preprocessing algorithm is not used an algorithm to discovery stemming basic words.

References

[1] T. Wijaya, "Perancangan Alat Ukur Indeks Usabilitas Pada Mesin Pencari (Search Engine)," Surakarta, 2011.
[2] K. Curran and J. M. Glinchey, "Vertical Search Engines," ITB Journal, vol. 8, no. 2, pp. 22-27, (2007).
[3] G. C. K. Yew, "Search engines and Internet defamation: Of publication and legal responsibility," Computer Law & Security Review, no. 35, pp. 330-343, 2019.
[4] K. Curran, Understanding the Internet, Oxford: Cando Publishing, 2009.
[5] Deepak and D. Mahal, "Migrating Parallel Web Crawler: A Review," International Journal of Innovations in Engineering and Technology (IJET), vol. VII, no. 3, pp. 312 - 320, June 2017.
[6] F. Ye, Z. Jing, Q. huang and Y. Chen, "The Research of a Lightweight Distributed Crawling," Computer Society, pp. 13-15, 2018.
[7] W. Yan and L. Pan, "Designing Focused Crawler Based On Improved," Tenth International Conference on Advanced Computational Intelligence (ICACI), pp. 319 - 323, 29-31 March 2018.
[8] Z. Shi, M. Shi and W. Lin, "The Implementation of Crawling News Page Based," Intl Conf on Applied Computing and Information Technology/3rd Intl Conf on Computational Science/Intelligence, pp. 348 - 351, 2016.
[9] M. Soulemane, M. Rafiuzzaman and H. Mahmud, "Crawling the Hidden Web: An Approach to Dynamic Web Indexing," International Journal of Computer Applications (0975 – 8887), vol. 55, no. 1, pp. 7 - 15, 2012.
[10] S. Sharma and P. Gupta, "The Anatomy of Web Crawlers," International Conference on Computing, Communication and Automation (ICCCA2015), pp. 849 - 853, 2015.
[11] G. H. Agre and N. V. Mahajan, "Keyword Focused Web Crawler," International Conference On Electronics And Communication, pp. 1089-1052, 2015.
[12] S. Rawat and R. P. D, "Efficient Focused Crawling based on Best First Search," International Advance Computing Conference (IACC), pp. 908-911, 2013.
[13] R. H. Rosman, "Implementasai Focused Crawler dengan menggunakan Content Similarity dan Structure Analysis," 2012.
[14] E. Uzun, "A Novel Web Scraping Approach Using the Additional Information Obtained From Web Pages," IEEE Acces, vol. 8, pp. 61726-61740, 13 April 2020.
[15] K. S. Shetty, S. Bhat and S. Singh, "Symbolic Verification of Web Crawler Functionality," International Conference on Computer Communication and Informatics (ICCCI -2012), pp. 1-6, 10 - 12 January 2012.
[16] S.-m. Kim and Y.-g. Ha, "Automated Discovery of Small Business Domain," BigComp, pp. 481-484, 18 January 2016.
[17] Y. D. Pramudita, S. S. Putro, E. M. Rochman, A. Y. Rofiqi, A. Jauhari, I. O. Suzanti and A. Rachmad, "Sports News Speak Indonesian Clustering Method Using Modified K-Medoid," in International Conference on Computer Science and Engineering Technology Universitas Muria Kudus, Surakarta, 2018.
[18] S. Ghankutkar, N. Sarkar, P. Gajbhiye, S. Yadav, D. Kalbande and N. Bakereywala, "Modelling Machine Learning For Analysing Crime News," in 2019 International Conference on Advances in Computing, Communication and Control (ICAC3), Mumbai, India, India, 2019.
[19] S.-H. Hong, S.-K. Lee and J.-H. Yu, "Automated management of green building material information using web crawling and ontology," Automation in Construction, no. 102, pp. 230 - 244, 05 March 2019.
[20] S. AlZu'bi, D. Aqel, A. Mughaidi and Y. Jararweh, "A Multi-levels Geo-location based Crawling Method for Social Media Platforms," *Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, pp. 494 - 498, 2019.