Vertical Search Engine for Academic Publications

A.S. YÜKSEL[1] and M.A. KARABIYIK[2]

[1] Suleyman Demirel University, Isparta/Turkey, asimyuksel@sdu.edu.tr
[2]Suleyman Demirel University, Isparta/Turkey, ma.karabiyik@gmail.com

***Abstract* – With the advancing technology, the storage of large amounts of data has become possible. Unstructured nature of data makes it difficult to access. Many sectors demand access to specific information within their area. Thus, it has emerged the concept of vertical search engine.**

In our study, a crawler was designed to filter reliable sites. The designed crawler only adds results related to academic publications to the database. Naive Bayes classifier algorithm was employed to identify the science branch of an academic publication by using its abstract. According to our experiments, the accuracy rate of developed vertical search engine was 70%. The application is designed in a way that it can self-learn so that the success rate can increase.

***Keywords* – Vertical Search Engine, Machine Learning, Naïve Bayes Classifier.**

I. Introduction

In the last twenty years, it has become easier for the corporate and individual users to have their own space in internet. For this reason, the increase in the number of websites has brought the search engines to attention. Search engines are systems that return results by entering keywords. They keep specific information and addresses of websites on their database. These databases are expanded by tools called crawlers.

Homonym words and similar terms used in different fields adversely affect search results. Users cannot get the results they want in a standard keyword search. Standard search engines have become insufficient over time. Insufficient search engines and developed artificial intelligence technologies have led to new ideas on search engines. Vertical search engines are products of these ideas [1]. The main features of the vertical search engines are the semantic review of the key text entered and the result pages being restricted, giving the user more specific result. Therefore, they can be applied to many areas [2].

In this study, academic publications were determined as subjects of the vertical search engine. The search results are returned to users through examining the abstracts of the publications. Developed search engine is composed of two modules. In the first module, science branch of the academic publication is identified. The second module contains the crawler that will collect the results that are to be returned to the user.

In the first part of our study, the text classification was done by applying supervised learning method. Supervised learning is a structure with input and output values [3]. In our experiments, Naïve bayes classifier and SVM classifiers were compared. The best-performing classifier was integrated into the system. Comparisons were made by using timing and accuracy rate criteria.

In the second part of our study, developed crawler identifies reliable websites. Journal, book and conference data were collected through these reliable websites. There are 44492 publications in the database collected by the crawler.

II. Related Works

Springer has developed a search engine to search through their own journals. It has approximately 2600 journals. Users can search using manuscript title, abstract or research field. Abstract and research field criteria are mandatory parameters for searching [4].

In the search engine developed by Elsevier, natural language processing techniques were employed. A wordlist has been created according to scientific fields. Topic detection was made by comparing created wordlist with the input text. This technique was named as fingerprinting. Search can be done using manuscript abstract, title or research field. The search engine returns results from the journals in Elsevier [5].

Enago has developed an application that searches within the open access journals indexed by Direct Access Journals (DOAJ). Only manuscript abstract is used as the search parameter. The results are shown with the percentage values that is called confidence index [6].

Edanz has developed a search engine with different search options. Users can apply filters such as journal name, publisher name, workspace and abstract. Additionally, Filters such as open access, impact factor, SCI index and SCI-e index can be applied. Edanz search engine returns more results since their database does not store information limited to a specific field as it is in other search engines [7].

III. Methodology

The application was developed as two modules. In the first module, text classification is done. The second module is designed as a crawler.

A. *Crawler*

Crawler is one of the important structures of search engines. Crawler is a program that autonomously browses web pages one by one. It takes the meaningful content of web pages and saves them to the database. Crawler repeats this process continuously [8].

For our application, most important factors that affects the results are consistency and clarity of data. Therefore, the crawler scans the reliable websites instead of random websites. Selecting reliable websites as starting point provides two advantages for application. These are smaller search space and higher performance.

The crawler collects data from browsed websites. This process has low performance when it is developed with standard methods. Therefore, we applied parallel programing techniques for high performance.

Another problem is code mistakes in browsed websites. Code mistakes complicate accessing data. We designed the crawler in a way that it is not affected from this kind of mistakes.

The crawler collected 44492 journals, conferences, books from the Internet. The problem that aroused while the data was collected was that the crawler could not access information due to website time outs. Loss of information due to timeouts was around 20%. After updating the system with parallel programming techniques, the data loss was reduced to a small rate of 0.003%.

For the vertical search engine to function properly, the results must be based on certain constraints. In this study, the constraints are provided by the use of previously mentioned reliable sites. Reliable sites accommodate the largest databases for academic publications. The following websites were considered as reliable.

- http://www.scimagojr.com/
- http://www.scijournal.org/
- http://mjl.clarivate.com/

28000 scientific publications were collected from these websites. Approximately 17000 of these publications are labelled with citation index and approximately 11000 of these publications are labelled with their impact factor values.

B. *Text Classification*

Machine learning methods were applied in the text classification process. The training set used in machine learning was designed in supervised manner. In the training set, the manuscript abstract is prepared as the input value and the scientific branch as the output value. Fourteen scientific branches were created as output values.

Abstracts were converted into vectors with bag of words (BoW) technique [9]. The classification was done by classifier algorithm on vectors. The best classifier algorithm was applied as a result of comparing Naïve Bayes classifier and Support Vector Machine (SVM) classifiers. Figure 1shows classification process.

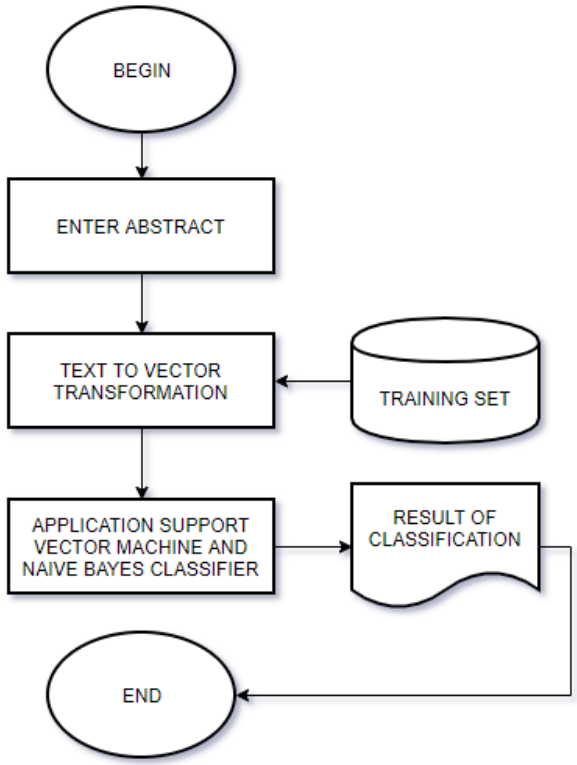

Figure 1: Classification process.

SVM, one of the machine learning algorithms, is a new algorithm based on statistical learning theory, which shows higher performance than traditional learning methods in solving classification problems such as pattern recognition and speech recognition [10].

The SVM algorithm draws lines to separate classes. These lines are called hyperplane. The purpose of these lines is to separate the classes from the boundary and to separate the classes from the boundary. If the test data to be classified is close to the line, it accepts the test data from that class.

Naive Bayes (NB) classifier algorithm is named after British mathematician Thomas Bayes. It is a statistical classifier and it can predict the possibility of belonging to a particular class. NB classifiers assume that the effect of a property value to a given class is independent of the values of other properties [11]. NB is a classifier algorithm that analyzes the relation between a set of values and other sets [12]. The formula for this method is shown in Equation 1.

$$P(A|B)=P(B|A)P(A)/P(B) \quad (1)$$

SVM and Naive Bayes classifier algorithms were compared by using time and accuracy metrics. The first comparison was performed over performance. Table 1 shows the accuracy comparison of SVM and SB for randomly choosen articles selected from Environmental Sciences.

Table 1: Accuracy results for Environmental Science.

Real Branch	SVM	SB
Environmental Science	Environmental Science	Veterinary
	Environmental Science	Environmental Science
	Social Sciences	Environmental Science
	Environmental Science	Environmental Science
	Environmental Science	Environmental Science
	Economics, Econometrics and Finance	Economics, Econometrics and Finance
	Veterinary	Environmental Science
	Environmental Science	Environmental Science
	Veterinary	Veterinary
	Agricultural and Biological Sciences	Environmental Science

In test process, tests have been performed with all science branches. The most complex results in terms of accuracy have been in environmental sciences. Therefore, the test were conducted for this science branch. Accuracy performance was almost equal in two classifiers. Table 2 shows the time and accuracy performances for randomly chosen articles by applying NB algorithm.

Table 2: Test results for NB.

Result	**Time (seconds)**
Environmental Science	4,631
Environmental Science	3,225
Social Sciences	3,447
Environmental Science	3,355
Energy	3,424
Environmental Science	3,309
Veterinary	3,299
Environmental Science	3,107
Environmental Science	3,341
Environmental Science	3,395
Success Rate: 70%	**Average: 3,453**

In both NB and SVM classifier algorithms, 10 abstracts were used in the tests. Table 3 shows the time and accuracy performances for SVM.

Table 3: Test results for SVM.

Result	**Time (seconds)**
Environmental Science	4,504
Environmental Science	4,472
Social Sciences	4,316
Environmental Science	4,307
Environmental Science	4,379
Environmental Science	4,228
Social Sciences	4,468
Environmental Science	5,354
Environmental Science	5,25
Agricultural and Biological Sciences	5,364
Success Rate: %70	**Average: 4,664**

It was observed that the number of data in the training set increased the accuracy performance. Therefore, the application allows users to add more data to the training set resulting the training set to be regularly updated. The computer that was used in experiments had Core i5 2.53 Ghz quad core processor, 4096 MB ram (1333 Mhz) and 250 GB SSD (540MB/s write, 520 MB/s read) hard drive. 70% success was achieved in the test with 140 text summaries. The time performance is about 3 seconds.

C. Application

A prototype is designed for this study. The main page is shown in Figure 2.

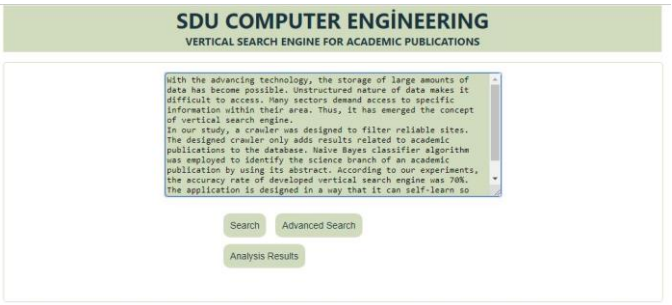

Figure 2: Main page of application.

In the main screen, there is a text box where a user can write manuscript abstract. This text box retrieves the text data to be translated into the feature vector. A search button for filterless search and an advanced search button for filtered search are presented to the users. Furthermore, classification results are shown to the user. Figure 3 shows the search page with advanced search button clicked.

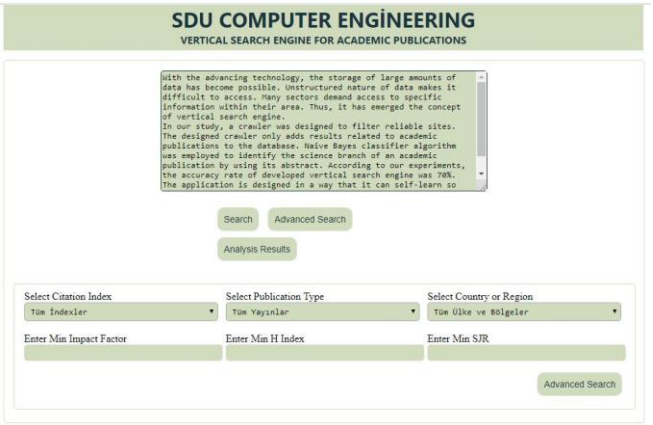

Figure 3: Search page with advanced options.

By clicking on advanced search button, six different filters appear. These are citation index, publication type, country and region, impact factor, H index and SJR point. The results are presented as in Figure 4.

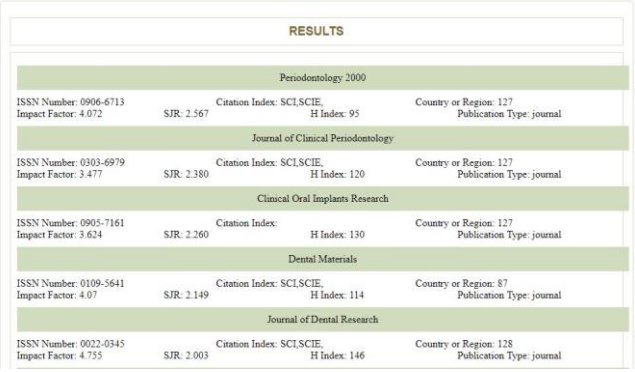

Figure 4: The results page.

Results are shown in result screen from sorted from the largest to the smallesst according to the impact factor. The publications in results are given with ISSN numbers for easy access. Figure 5 shows the analysis result page.

Figure 5: Analysis result page.

The proportional distribution of the results of the abstract entered in the analysis results screen is also presented. Users can update the result if they think the search result is not accurate.

IV. Conclusions

In this study, a vertical search engine has been developed for academic publications. The academic publications were collected from reliable websites via developed crawler. In the text classification process based on machine learning, Naïve Bayes classifier algorithm was employed. 70% classification accuracy rate was achieved. Self-learning feature of the system allows users to train the system for more accurate results.

References

[1] C. D. Manning, P. Raghavan, and H. Schütze, *An Introduction to Information Retrieval*, no. c. England: Cambridge University Press, 2009.

[2] C. Razbonyalı, "Research on Vertical Search Engines and Devoloping an Application on Vertical Search Engine," Trakya University, 2011.

[3] S. Kulkarni and M. Mushrif, "Notice of Violation of IEEE Publication Principles Comparative Study among Different Neural Net Learning Algorithms Applied to Rainfall Predication," *2014 Int. Conf. Electron. Syst. Signal Process. Comput. Technol.*, no. April 2008, pp. 209–216, 2014.

[4] Springer, "No Title," 2004. [Online]. Available: https://journalsuggester.springer.com/. [Accessed: 25-Oct-2017].

[5] Elsevier, "No Title," 2012. [Online]. Available: https://journalfinder.elsevier.com. [Accessed: 25-Oct-2017].

[6] Enago, "No Title," 2005. [Online]. Available: https://www.enago.com/academy/journalfinder/. [Accessed: 27-Apr-2018].

[7] Edanz, "No Title," 2000. [Online]. Available: https://www.edanzediting.com/journal-selector. [Accessed: 27-Apr-2018].

[8] X. Ma, ChaoSong, MeinaXu, KeZhang, "Web Service discovery research and implementation based on semantic search engine," *2010 IEEE 2nd Symp. Web Soc.*, pp. 672–677, 2010.

[9] L. Tian and S. Wang, "Improved Bag-of-Words Model for Person Re-identification," vol. 23, no. 2, pp. 145–156, 2018.

[10] L. Oneto *et al.*, "Dynamic delay predictions for large-scale railway networks: Deep and shallow extreme learning machines tuned via thresholdout," *IEEE Trans. Syst. Man, Cybern. Syst.*, vol. 47, no. 10, pp. 2754–2767, 2017.

[11] E. S. İ. N and Y. K. Çelİ, "Veri Madencili ğ inde Kay ı p Veriler İç in Kullan ı lan Y ö ntemlerin Kar şı la ş t ırı lma sı."

[12] H. H. M. A. A.-R. Al-Hudairy and U. of Louisville, "Data mining and decision making support in the governmental sector," Lousville University, 2004.