International Journal of Engineering & Technology, 7 (3.20) (2018) 154-157

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Implementation of Vector Space Model in Online Jobs Vacancy Aggregator

Tjut Awaliyah Zuraiyah , Fajar Delli Wihartiko and Edwin Effendi

Departemen of Computer Science Pakuan University
Corresponding Author Email: tjut.awaliyah@unpak.ac.id

Abstract

Job vacancy aggregator is a system that facilitates users in finding the desired job vacancy, especially in the field of information technology. Job vacancy data collected from various job sites such as http://id.jobsdb.com, http://www.jobs.id, http://www.monster.co.id and http://www.jobstreet.co.id using web scraping techniques to extract job vacancy data that is stored in the HTML structure. The collected data is then processed to facilitate the retrieval concept by vector space model method, by using vector space model data which is found to be sorted based on the similarity level between the query which is typed by the user with the job vacancy data is stored in the database. In addition system can also perform email jobs sent via email to registered users. With the development of an online job vacancy aggregator, it can be used as a media job vacancy information, especially in the field of information technology (IT).

Keywords: Vector Space Model, Online Job Vacancy, Scapping

1. Introduction

Along with the times, technology in the world is progressing very rapidly, one of them in the field of Internet technology with a very large quantity of websites. Similarly, the website regarding job vacancies almost every time there are always new jobs. In order to anticipate the traffic of job vacancy information on the internet then develops collecting technology (aggregator) and job seekers, especially in the field of information technology. The system collects job vacancy data from various job sites such as id.jobsdb.com, www.jobs.id, www.monster.co.id and www.jobstreet.co.id. Using the web scraping technique for extracting the job vacancy data that stored in the HTML structure of job sites. The system is made using the concept of information retrieval, using the vector space model to calculate the similarity beetwen query input by user and job vacancy data stored in the database, In addition system can also perform email recomended jobs sent via email to registered users.
With the construction of this online job vacancy aggregator hope it can help users in finding the desired job without having to open a number of job sites and can be used as a media of job vacancy information especially in the field of information technology (IT).

2. Material and Methods

Studies that have been done with regards to web scraping and aggregator, first research by [1], which discusses the development of web aggregator Bogor Institute of Agriculture uses a single stream aggregation. The second research was conducted by [7] which discusses the application of Indonesian news aggregator based on android supported by the recommendation system. And the third conducted by [4] which discusses the design and manufacture of job search engines using the cosine similarity.

Flowmap a graphical depiction of the steps and the sequence of procedures of a program. Flowmap analysis and programmers to help solve the problem into segments that are smaller. Flowmap system to be created can be seen in Figure 1.

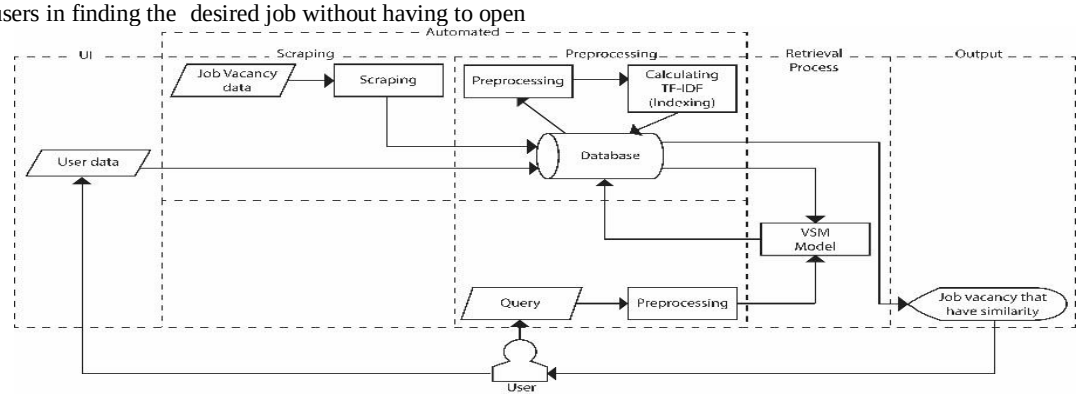

Figure 1. Flowmap system to be created.

2.1 UI (User Interface)

User Interface acts as a medium of communication between the user and the system, in this case the UI receives data in the form of data users who wish to enrol in the system, one of which is the e-mail later it will be used as a medium for distributing job information recommended by the system based query ever searched by users.

2.2 Scraping

Scraping flowmap explain how the workflow in the data collection system on the job site. Flowmap scraping can be seen in Figure 2.

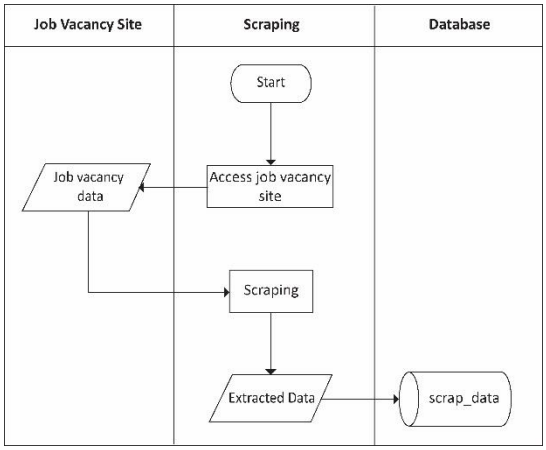

Figure 2: Flow map scraping.

Flow map scraping begins by accessing the job sites that serve as a source of data, which previously has been created for scraping template for each site based on the HTML tags that flank the desired data. Template scraping for id.jobsdb.com sites shown in Table 1.

Table 1. Scraping Template JobsDB site

Num	Attribute	Tag Name
1	Job Title	
2	Link Address	
3	Company logo	<Img src=></img>
4	Company name	
5	Company Location	
6	Short description	

2.3 Pre-processing

Design flow map pre-processing explain how where the workflow system in normalizing the data that has been in the scrap from job sites, so the data is ready to do the retrieval process. Flow map pre-processing can be seen in Figure 3.

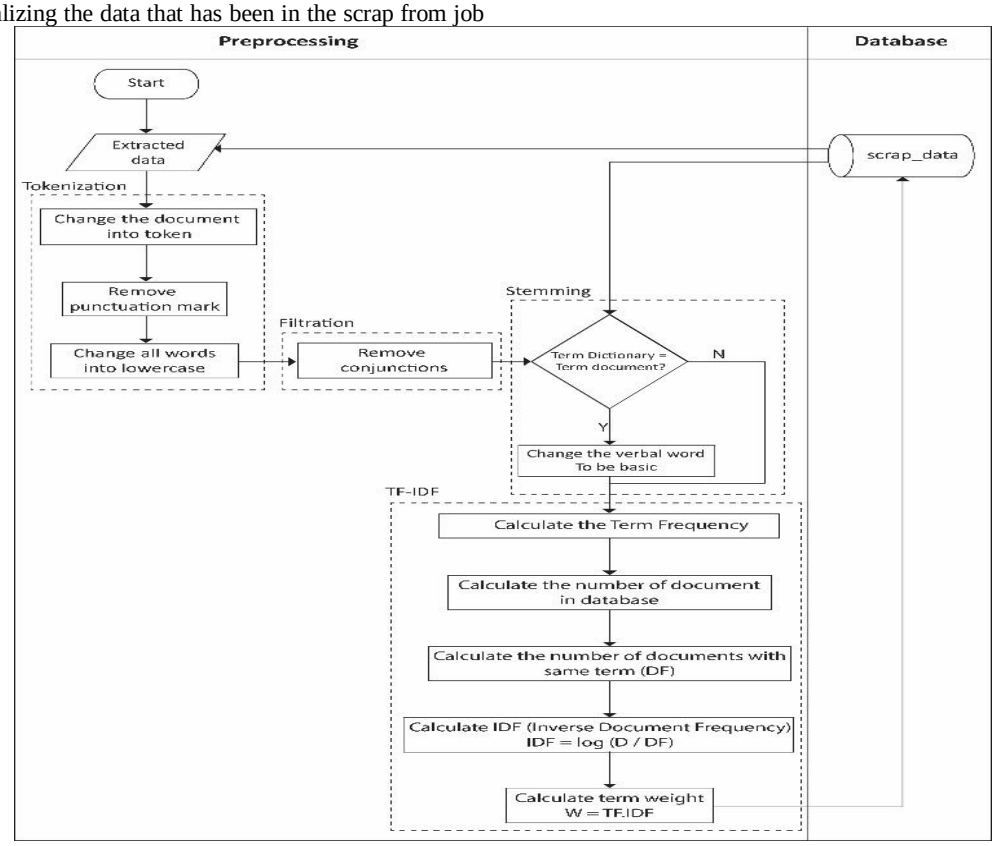

Figure 3. Stages of Preprocessing.

The result data preprocessing then entered back into the database to be processed at a later stage. Here is a simple example of preprocessing

D1: "IT Support"
a. Tokenization: it, support.
b. Filtration: it, support.
c. Stemming: it, support.

D2: "Desktop Engineer"
a. Tokenization: desktop, engineer.
b. Filtration: desktop, engineer.
c. Stemming: desktop, engineer.

D3: "Desktop - Printer Support Engineer Jakarta"
a. Tokenization: desktops, printers, support, engineer, jakarta.
b. Filtration: desktops, printers, support, engineer, jakarta.
c. Stemming: desktops, printers, support, engineer, jakarta.

The results of the preprocessing stage will proceed to the TF-IDF calculation phase shown in Table 2.

Table 2. Calculation of TF-IDF

Terms	Q	D1	D2	D3	DF	D / df	IDF	Weight			
								Q	D1	D2	D3
It	0	1	0	0	1	3/1 = 3	0.4771	0	0.4771	0	0
Desktop	1	0	1	1	2	3/2 = 1.5	0.176	0.176	0	0.176	0.176
Support	0	1	0	1	2	3/2 = 1.5	0.176	0	0.176	0	0.176
Engineer	1	0	1	1	2	3/2 = 1.5	0.176	0.176	0	0.176	0.176
Printer	0	0	0	1	1	3/1 = 3	0.4771	0	0	0	0.4771
Jakarta	0	0	0	1	1	3/1 = 3	0.4771	0	0	0	0.4771

2.4 Design Flow map Retrieval

Design flow map retrieval explain how where the workflow system to surf the job back to the data stored in the database. Flow map retrieval can be seen in Figure 4.

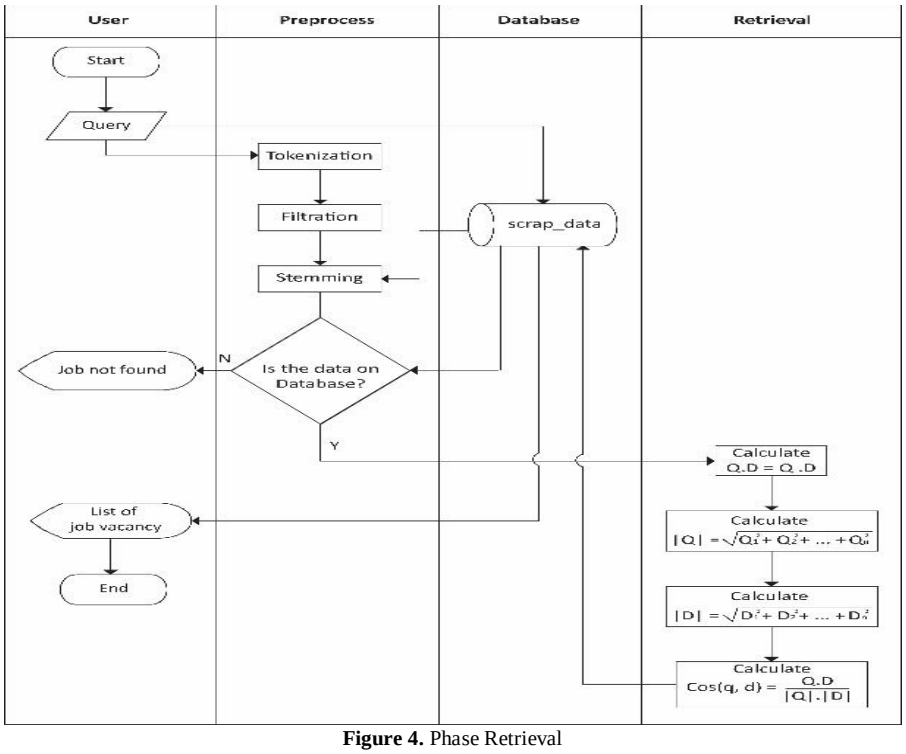

Figure 4. Phase Retrieval

Flow map retrieval begins from user activities that do a search, a query of the user entered into the database to be reprocessed as data to perform a job recommendation, while the query was done pre-processing phase consisting of tokenization, filtration, and stemming.
Results of query pre-processing stage will be matched with job vacancies data set in the database using the Vector Space Model with the following formula.

$$R\,(Q.D) = \cos\theta = \frac{Q \cdot D}{|Q||D|}$$

Where:
Q = Weight of query D = Weight of document
| Q | = Length of query | D | = Length of document

then the calculation result will be entered into the database and displayed to the user in the form of a sequence based on the level of similarity between the query with the data collection job in the database. Here are the steps retrieval using the vector space model calculations using the data in the Table 2. Calculation of TF-IDF.

1. Multiply the weight between the weight of the query term weights on each document
   Q. D1 = 0.176 * 0 + 0.176 * 0 = 0
   Q. D2 = 0.176 * 0.176 + 0.176 * 0.176 = 0,0611
   Q. D3 = 0.176 * 0.176 + 0.176 * 0.176 = 0,0611

2. Calculate the length of the query.
   $| Q | = \sqrt{0.176^2 + 0.176^2} = 0.2490$

3. Calculates the length of each document.
   $| D1 | = \sqrt{0.4771^2 + 0.176^2} = 0.5085$
   $| D2 | = \sqrt{0.176^2 + 0.176^2} = 0.2490$
   $| D3 | = \sqrt{0.176^2 + 0.176^2 + 0.176^2 + 0.4771^2 + 0.4771^2} = 0.7404$

4. Calculate the cosine similarity by dividing the weight of QD_n by the multiplication of the long query ($| Q |$) and the length of the document ($| D_n |$)
   $\text{Cos}(Q, D1) = \frac{0}{(0.2490 * 0.5085)} = 0$
   $\text{Cos}(Q, D2) = \frac{0,0611}{(0.2490 * 0.2490)} = 0.9855$
   $\text{Cos}(Q, D2) = \frac{0,0611}{(0.2490 * 0.7404)} = 0.3314$

   Similarity level = D2, D3, D1.

3. Result

The search feature job listings using the concept of information retrieval (IR), the advantages of the concept of Information Retrieval is the use of the principle of relevancy in search results with a range between 0 (irrelevant) to 1 (relevant), when compared with searches using SQL commands can be seen clearly the difference is that the search results only recognize the value of 0 (not found) and 1 (found) that the range of the data found fewer. Differences in search results between Information Retrieval and SQL commands can be seen in Figures 5 and 6 where the job data found by the IR system is 9 data, whereas the SQL command finds only 2 data.

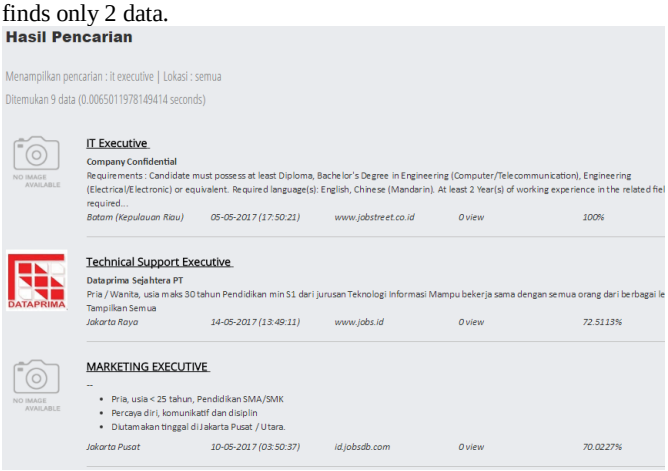

Figure 5. Results with Information Retrieval Concepts

Figure 6. Results with SQL Commands

4. Discusion

Online job vacancy aggregator is designed to enable job seekers in finding a job, especially in the field of Information Technology (IT) without having to open a number of job sites, so as to shorten the time to find a job which are desired.

Using the concept of information retrieval where this concept will provide search results that are relevant to a range of data found to be more widespread, unlike the usual SQL searches only find data that is matched records. Searches using the vector space model to measure the degree of similarity between the query given a job with the data in the database.

The system has been through a pilot phase, including structural testing, functional trials, and validation trials. Validation test is done by:

1. Query combination testing that has been conducted shows that the order of the word order has no influence on the results, because the method of vector space model does not account for the sort order of words, things that affect the search results that the weight of the search terms. The testing shown in Appendix 1.
2. Calculation of the precision and recall, after testing for 5 times with a different query obtained an average of 74.52% system precision and recall of 100%, which means that the system can provide data relevant with the search level Precision of 74.52%, the search result and precision recall trial are shown in Appendix 2 and 3.

From the test results, it can be concluded that the validation test has been successful and the aggregator of this job vacancy is feasible to use.

References

[1] Asry, H. F. 2011. Pengembangan Web Aggregator Institut Petanian Bogor Menggunakan Single Stream Aggregation. Skripsi. Fakultas Matematika dan Pengetahuan Alam IPB, Bogor.

[2] Julian, L. R. and Natalia, F. 2015. The Use of Web Scraping in Computer Parts and Assembly Price Comparison. Departemen Sistem Informasi Universitas Multimedia Nusantara, Banten.

[3] Karmayasa, O. and Mahendra, I. B. 2012. Implementasi Vector Space Model dan Beberapa Notasi Metode Term Frequency Inverse Document Frequency (TF-IDF) pada Sistem Temu Kembali Informasi. Program Studi Teknik Informatika Universitas Udayana, Bali.

[4] Raharjo, D. S., Solihin, F. and Santosa, I. 2015. Perancangan dan Pembuatan Mesin Pencari Lowongan Pekerjaan Menggunakan Metode Cosine Similarity. Program Studi Teknik Informatika Universitas Trunojoyo, Madura.

[5] Singh, V. K. and Singh, V. K. 2015. Vector Space Model: An Information Retrieval System.

[6] Vijayarani, S., Ilamathi, J. and Nithya. 2015. Preprocessing Techniques for Text Mining.

[7] Wahono, N. V., Wibowo, A and Intan, R. 2014. Aplikasi Indonesian News Aggregator Berbasis Android yang Didukung oleh Sistem Perekomendasi. Program Studi Teknik Informatika Fakultas Teknologi Industri Universitas Kristen Petra, Surabaya.