JOB MATCHING PADA DATA I-CDC MENGGUNAKAN LATENT SEMANTIC ANALYSIS JOB MATCHING ON I-CDC DATA USING LATENT SEMANTIC ANALYSIS

J. Catur Prasetiawan[1]

[1]Magister Teknik Informatika, Fakultas Teknik Informatika, Universitas Telkom

Abstrak

Pencarian kandidat terbaik untuk suatu lowongan pekerjaan bukanlah sesuatu yang sederhana. Memerlukan banyak proses yang melibatkan banyak orang serta membutuhkan waktu. Sistem job matching yang terkomputerisasi bisa menjadi solusi agar masalah ini menjadi lebih sederhana. Konsep dari document matching bisa diterapkan dalam persoalan ini. Antara dokumen pencari kerja dan dokumen lowongan kerja akan dicocokkan berdasarkan kemiripannya. Pada penelitian ini terbukti ada cara yang ditemukan untuk menangani data numerik pada LSA. LSA digabungkan dengan selection approach mampu memberikan presisi yang lebih baik. Penelitian ini juga membuktikan bahwa langkah pre-processing (stopping/menambahkan stop word) mempengaruhi hasil. Pre-processing dapat menaikkan presisi dari sistem dengan LSA yang dimodifikasi untuk numeric (pendekatan seleksi & pembobotan). Dapat ditunjukkan juga bahwa rank k approximation yang optimal dapat dicapai pada paruh pertama dari rentang k, untuk data pelamar sebanyak 100 yang dibandingkan dengan data lowongan sebanyak 20.

Kata Kunci : , Stopping Step, Stop Word, Numeric Data, Rank Approximation

Abstract

Finding the best candidates for a job vacancy is still not a simple thing. It takes multiple process involving so many person, and also takes time. Building a computerized job matching system might be a solution to make this problem simpler. The concept of document matching could be implemented for this case. Between the document of job seekers and the document of job vacancy will be matched by looking their similarity. It was proved that there is a way to handle numeric data in LSA. LSA collaborated with selection approach for numeric is able to improve the precision .This study also proved that pre-processing step (stopping step/ adding stop word) influence the result. Pre-processing step could improve the precision of system with modified LSA for numeric (with selection & weighting approach). This study also showed that the optimal rank k approximation can be obtained at the half first range of k, for 100 numbers data of job seekers compared with 20 numbers data of vacancy.

Keywords : LSA, Stopping Step, Stop Word, Numeric Data, Rank Approximation

Chapter 1. The Problem

This study is dealing with job matching on Infocom Career Development Center (i-CDC) IT Telkom. Before it is discussed in great detail, The Rationale, Theoretical Framework, Conceptual Framework/Paradigm, Statement of the Problem, Hypotheses, Assumption, Scope and Delimitation, Importance of the Study, and Definition of terms are discussed briefly in this chapter.

1.1. Rationale

Unemployment is an important issue in developing countries like Indonesia. According to BPS (Central Bureau of Statistics) the number of high educated unemployed tends to increase from 585,358 in 2004 to 1,153,350 in 2010. This issue could cause social problems if not managed properly. To handle this issue, the government has required universities to ensure graduates will not be unemployed for long. Waiting time of graduates for the job has been one of the requirements for accreditation of study programs.

To respond to this regulation IT Telkom, as a private university, has set up i-CDC (infocom career development center) as a bridge between graduates with industries. The i-CDC collect and provide information about job vacancies to the graduates. i-CDC also offers the qualified graduates to the industries. Until now, i-CDC has had 2611 members and collected about 100 job vacancies information per month. All data in the form of text documents.

This study builds job matching system using the concept of document matching. Between the documents of graduate's CV and job vacancy will be matched by looking their similarity. This method is also well known as document matching. This study intends to provide optimal way to find the similarity between documents by using Latent Semantic Analysis (LSA) approach.

A Study by Cheng Kam Ching in 2011 [3] on the job matching has shown that LSA method combined with TFIDF (Terms Frequency Inverse Document Frequency) weighting (LSA-TFIDF) can improve the accuracy of the original LSA approach. The LSA-TFIDF approach has 63.7% accuracy, meanwhile the original LSA has 58.4%. Previous study did’nt

explore the way to compare numerical data. This study will look for the effect of numerical data processing on the accuracy of original LSA.

1.2. Theoretical Framework

This study implement the concept of text retrieval for job matching on i-CDC data. Firstly, the system will extract the important terms from documents. Then, it will build a matrix based on the extracted terms, the columns contain the terms and the rows represent the documents. Singular Value Decomposition (SVD) is used to decompose the matrix in order to analyze the terms and it's contribution for each documents. This method is well known as 'Latent Semantic Analysis' (LSA) that usually implemented in text retrieval system.

LSA is chosen because it is a concept-based approach rather than keywords matching which is prone to failure. Thomas K Landauer said," *LSA is a theory and method for extracting and representing the contextual usage meaning of words by statistical computations applied to a large corpus of text".*[6]

1.3. Conceptual Framework/Paradigm

The research variables and the relationship with the conceptual research will be discussed in this sub chapter.

Table 1-1 Research Variable

No	Relationship with Conceptual Research	Variable
1	Size of matrix, time needed to run the program	Number of terms
2	Accuracy of similarity	Rank of approximation
3	Precision	Relevant documents

1.4. Statement of the Problem

Statement of the problems of this study are:

1. How to compare numeric data in LSA ?
2. What are the correlations of pre-processing step and precision of system?
3. What is the optimal rank approximation influences the precision of system?

1.5. Hypotheses

The hypotheses of this study are:

1. There is a way to compare numeric data in LSA.
2. There are correlation of pre-procesing and precision.
3. There is an optimal rank approximation influences the precision.

1.6. Assumption

The assumptions of this study are :

1. The data of job seekers and the data of job vacancies are written in standard format
2. All data does not contain so many useless information

1.7. Scope and Delimitations

The scopes and delimitations are :

1. This study is conducted based on the data taken from i-CDC of IT Telkom
2. This study used the data of job vacancies written in Bahasa Indonesia on 2010
3. This study used the data of job seekers written in Bahasa Indonesia on 2010
4. This study compared job seekers and job vacancies data based on hardskills required.

1.8. Importance of the Study

This study is important to build a system so that i-CDC can recommend qualified graduates/ job seekers proactively and quickly.

1.9. Definition of Terms

LSA	Latent Semantic Analysis, one of text retrieval algorithm. Also well known as Latent Semantic Indexing (LSI).
Terms	The words can be extracted from a documents
SVD	Singular Value Decomposition, one of method to decompose a matrix.
R-precision	Metric to evaluate effectiveness of the job matching system
Rank *k* Approximation	The largest k singular value will be retained.

Chapter 5. Conclusions and Recommendations

Based on experiments that have been conducted, obtained the following conclusions and recommendations.

5.1. Conclusions

1. Stopping step (pre-processing) could improve the precision of system in case the use of LSA-selection method and LSA-weighting method. But in the use of Original LSA a slightly better result obtained without stopping step.
2. Handling numeric data by using LSA-selection approach combined with stopping step give the best precision of result. By adding stopping step, LSA-selection approach give 47.06% precision, LSA-weighting approach give 46.57% precision, meanwhile The Original LSA 45.45%. The use of numerical handling approach depend on the way of candidates selection. Selection by absolute criteria, everybody should pass all requirements, more suitable to do by using LSA-selection approach.
3. The optimal rank approximation (k) can be obtained at half first of range of k for 100 numbers of job seekers data to be compared with 20 number of job vacancy.

5.2. Recommendations

1. Exploring another numerical data handling method to obtain the better precision of system
2. Combining the LSA with another approach might be useful to improve the precision

Bibliography

[1] Aji, Rizki Bayu. 2011. Automatic Essay Grading System Using Latent Semantic Analysis Method.

[2] Baker, Kirk. 2005. *Singular Value Decompostion Tutorial*. http://www.cs.wits.ac.za/~michael/SVDTut.pdf. Downloaded on June 22nd 2011

[3] Ching, Cheng Kam., 2011, *Development of Job Matching Algorithm with Collective Learning Methods*,

[4] Garcia, Dr. Edel, 2006, *SVD and LSI Tutorial 4: Latent Semantic Indexing (LSI) How-to Calculations*, Mi lslita.com, **http://www.miislita.com/information-retrieval-tutorial/svd-lsi-tutorial-4-lsi-how-to-calculations.html**. downloaded on June 22nd 2011

[5] Garcia, Dr. Edel. 2006. *Singular Value Decompostion A Fast Track Tutorial*. http://www.miislita.com/information-retrieval-tutorial/singular-value-decomposition-fast-track-tutorial.pdf. Downloaded on December 30th 2010

[6] Landauer , Thomas K ., Peter W. Foltz, and Darrell Laham. 1998. *Introduction to Latent Semantic Analysis*. Discourse Processes, 25, 259-284. http://lsa.colorado.edu/papers/dp1.LSAintro.pdf. Downloaded on December 3th 2010

[7] Manning, Christopher D., Raghavan Prabhakar, Schutze Hinrich, 2008, *Introduction to Information Retrieval*, Cambridge University Press

[8] Neto, Joel Larocca, 2000, *Document Clustering and Text Summarization* , Pontificia Universidade Catolica do Parana