The 6th International Seminar on Science and Technology (ISST) 2020
July 25th, 2020, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia

Job Standard Parameters from Online Job Vacancy

Herlambang Haryo Putro and Nur Aini Rakhmawati
Departement of Informations System, Institut Teknologi Sepuluh Nopember, Surabaya
e-mail: herlambangharyo@gmail.com

***Abstrak*—The internet has provided an efficient and economical way to provide job vacancy information to applicants in a way that is far more dynamic and consistent than happened in the past. This is directly proportional to the number of companies exploiting online technology (job portals, company websites, etc.) to make job advertisements reach a growing audience. To further optimize the selection process concerning processing time and accuracy, researchers have begun developing a sophisticated search engine to automatically sort resumes based on job offer requirements. Overcoming this problem requires the Information Extraction (IE) process. Research on IE regarding vacancies actually already exists and has been applied. Research on IE regarding vacancies actually already exists and has been applied In addition, we want to develop previous research involving all job vacancies websites to a wider extent. We realize that this limitation makes the extraction process difficult. we try to define job problems more easily, such as identifying various JSPs on job vacancy websites in Indonesia as a first step in the extraction process. Our findings in the survey of 14 job vacancy websites are 26 job standard parameters including structure collection and extraction methods. This study provides a detailed description of each component of information extraction on the job vacancy website in Indonesia. Starting from identifying the type of Structured Extraction to the output of extraction. This study also developed JSP from previous research.**

Keywords—Job Standard Parameters, Entity, Method Extraction.

I. INTRODUCTION

THE internet has provided an efficient and economical way to provide job vacancy information to applicants in a way that is far more dynamic and consistent than happened in the past [1]. This is directly proportional to the number of companies exploiting online technology (job portals, company websites, etc.) to make job advertisements reach a growing audience [2]. However, this advantage can create a burden for recruiters, who have to sort out the number of resumes and curriculum vitae received, often expressed in various languages and formats. Similarly, job seekers spend a lot of time filtering job offers and restructuring their resumes to effectively communicate their strong points and meet job requirements. As a result, recruiters and job seekers often use a variety of special-purpose tools, such as job aggregators (including jobstreet.com and indeed.com) and social networks (including linkedin.com).

To further optimize the selection process concerning processing time and accuracy, researchers have begun developing a sophisticated search engine to automatically sort resumes based on job offer requirements [3]. These approaches can exploit, among others, supervised and unsupervised machine learning and even ontology. However, creating such a tool is a complex task that requires the identification of entities from Job Standard Parameters (JSP) [3] that affect the user's final choice.

Overcoming this problem requires the Information Extraction (IE) process. Information extraction process is used to extract structured content in the form of entities, relations, facts, terms, and other types of information that helps the data analysis [4]. Research on IE regarding vacancies actually already exists and has been applied [3],[5]. In addition, we want to develop previous research involving all job vacancies websites to a wider extent. We realize that this limitation makes the extraction process difficult. As data scientists, we try to define job problems more easily, such as identifying various JSPs on job vacancy websites in Indonesia as a first step in the extraction process.

So, our research questions are (1) How many JSPs can be defined on the job vacancy website in Indonesia? (2) What is the extraction structure? (3) How is the extraction methodology from each JSP? (4) How can we extract the extracted results? Based on observations about JSP, we believe that JSP can be further developed to create an appropriate two-way communication between the company and applicants.

II. METHOD

We divide the research methodology into 3 parts: Website Selection, Data Collection, and Data Analysis. each of which will be discussed in detail in the following subsections. See Figure 1.

A. Website Selection

The website criteria we are looking for are based on the availability of job vacancies in Indonesia. We use "Lowongan Pekerjaan" as a keyword for the first step. The next step is to choose websites that contain Bahasa Indonesia. Websites that meet the criteria are: glints.com, id.indeed.com, id.jobsdb.com, id.jooble.org, jobs.id, jobstreet.co.id, karir.com, karirhub.kemnaker.go.id, karirpad.com, lokerjogja.id, mamikos.com, terminalhrd.com, topkarir.com, urbanhire.com.

B. Data Collection

Data collection aims to identify JSP available on the website. One website accesses 20 pages randomly with different companies. This screening intends to look for sentence patterns in each JSP. The available JSPs are collected using Google Form in the same website labeling.

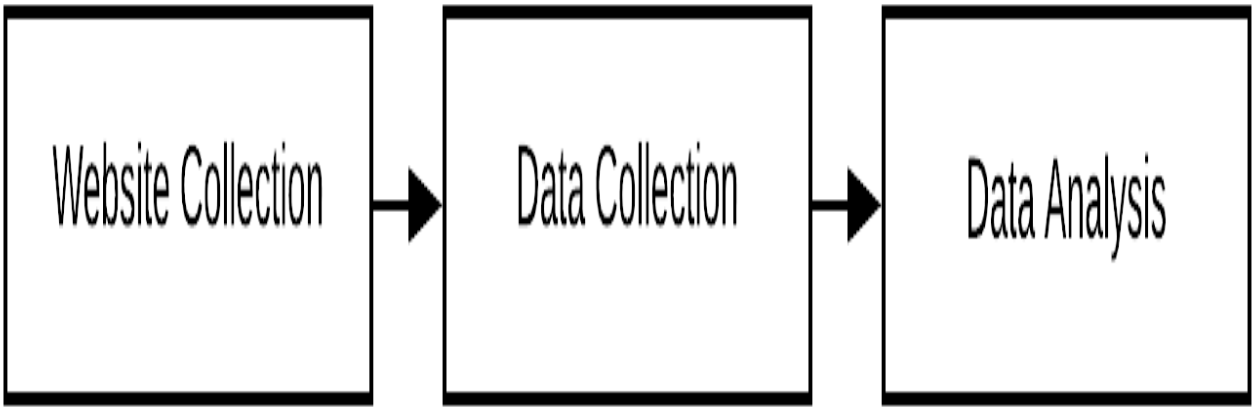

Figure 1. Methodology.

This will help identify how JSP patterns are displayed on the website.

C. Data Collection

Analyzing JSP data is assisted with the Google form feature. The identified JSPs are collected and categorized based on their purposed and Sarawagi's research [6]. This research includes any Type of Structured Extracted, Method used for Extraction, and Output expectations that should be stored in a database.

III. RESULT AND DISCUSSION

In this section, we discuss the detailed results of the data analysis. These details include the list of JSPs that exist in Indonesia. Then the Entity and Relationship in the JSP. As well as we discuss how JSP is extracted. And finally, how the extraction results.

A. Job Standard Parameter

Job Standards Parameters found can be categorized into 3 parts: Company, Job, and Job Description. Data about the company is summarized in the company section. The data consists of Company, Industry, Address, City, Telephone, Website, and Coordinates. Jobs consist of Title, Position, Domain, Salary, Placement, Employment Type. While job descriptions are identified as supporting data consisting of Requirements, Education, Major, Experiences, Soft-skills, Hard-skills, Age, Gender, Marital Status, Job Desk, Benefit, Opening Date, and Closed Date.

Company: Company name data. An interesting finding on jobs.id, The form of a company that is generally written on the prefix of the company name becomes the company suffix. For example "Indonesian Factory PT" which should be "PT Pabrik Indonesia". Industry: Data on the company's industrial fields. Address & City, not all job vacancy websites provide both of these data. However, address data can identify a company's city. Then, Phone, Website, and Coordinate are complementary data from a company. See the list in Table 1. Title: job name data. Some job names can identify Position & Placement. For example, "Supervisor Production Manufacturing Area Malang". "Supervisor" can be identified as a Position. While "Malang" can be identified as a Placement. Domain: occupational data, Salary: salary data which can be in the form of minimum, maximum, and even range. But other data can be information that the company does not want to mention salary. Employment Type: generally contain Full-time, Contract, and Internship information. Some websites even display semantic information. See the list in Table 2.

Requirements : Requirements: are qualifications of a job consisting of Education, Majors, Experiences, Soft Skills, Hard Skills, Age, Gender, and Marital Status. See the list in Table 3.

B. Type of Structure Extracted

Entities in this JSP are categorized as follows: Company, Industry, Address, City, Phone, Website, Coordinate, Title, Position, Domain, Employment Type, Education, Marital Status, Job Desk, and Benefit as a single value. While Salary, Placement, Major, Soft-skill, and Hard-skill as a multiple value. The type of relationship structured extracted is categorized as follows: Requirement.

The mistake of job vacancy provider websites is the lack of diversity in JSP on their services. Often several JSPs are considered as a Requirement. Though they can define it in more detail so that the accuracy of the needs between the company and applicants. So that the requirements are multi-entity as relationship structured extracted.

C. The Method Used for Extraction

We believe that the readers of this paper know the same field. So that each of them has a preference for how the extraction results are changed by the data type. However, this interesting thing lies in the relationship between Position and Experience. We identify some websites that display Experience with the same meaning as Position. Then, we also identified the strange behavior of JSP. We assume that the company has detailed candidate criteria. However, not all websites have the same features. Which finally forced all of the criteria to be placed in the Requirements.A hand-coded system requires human experts to define rules or regular expressions or program snippets for performing the extraction X6 Hand-coded methodology is the easiest methodology that generally extracts single-entity including the JSP entity that we have defined. However, several things need to be noted, among others, namely the Company, a small rule is required regarding the writing of the prefix form of the company.

Table 1.
Job Standard Parameter for Company

No.	Category	Job Standard Parameter
1	Company	Company
2.	Company	Industry
3.	Company	Address
4.	Company	City
5.	Company	Phone
6.	Company	Website
7.	Company	Coordinate

Table 2.
Job Standard Parameter for Jobs

No.	Category	Job Standard Parameter
1	Jobs	Title
2.	Jobs	Domain
3.	Jobs	Position
4.	Jobs	Salary
5.	Jobs	Placement
6.	Jobs	Employment Type

Table 3.
Job Standard Parameter for Jobs Descriptions

No.	Category	Job Standard Parameter
1	Jobs Description	Requirement
2.	Jobs Description	Education
3.	Jobs Description	Major
4.	Jobs Description	Experiences
5.	Jobs Description	Soft Skills
6.	Jobs Description	Hard Skills
7.	Jobs Description	Age
8.	Jobs Description	Gender
9.	Jobs Description	Marital Status
10.	Jobs Description	Job Desk
11.	Jobs Description	Benefit
12.	Jobs Description	Opening Date
13.	Jobs Description	Closed Date

Coordinate can be better defined with CoodinateX and CoodinateY which record the location of the company. On some websites, City is displayed with a complete location with the province and country. In contrast to the Placement, some displayed directly the destination city of the vacancy. The rules of Salary and Age have the same similarity. which is a minimum, maximum, and range. Interestingly, the description of the values of Salary and Age can be in the form of semantic sentences. Same as Experience, but Experience is not defined as a range value. Employment type and Education only have a few values. Employment type can be in the form of "Full-time", "Part-time", "Temporary", "Internship", even recently "Remote" is also included. Whereas Education can be in the form of "SMA / SMK", "Diploma", "Sarjana", and even can also be in the form of abbreviations D3, S1, S2, and S3. Gender and Marital status have fewer values. Gender can be "Pria", "Wanita", "Pria dan Wanita". Whereas Marital Status can be a "Lajang" and "Menikah". See the list in Table 4. Learning-based systems require manually labeled unstructured examples to train machine learning models of extraction [6]. The learning-based methodology will be used to extract and identify Major, Soft-Skill, and Hard-skill. We could not identify how many values from each JSP. So, it requires labeling for a better result.

The rule based method use several general rules instead of dictionary to extract information from text [7].The Rule based systems have been mostly used in information extraction from semi-structured web page. A usual method is to learn syntactic/semantic constraints with delimiters that bound the text to be extracted, that is to learn rules for boundaries of the target text. The rule-based methodology can extracts multi-entity and has a slightly more complex rule than hand-coded. An example of identifying City in Address is

({Orthography type = String}):Gedung ? ({Orthography type = String}):Street ({String = ","})({Orthography type = String}):City →Address = :Gedung & Street, City =:City
Another example of identifying Position, Placement, Employment Type in Title is ({Orthography type = String}):Position ? (Orthography type = String):Title ({String = "Area"})? ({Orthography type = String}):Placement? ({Orthography type = String}):Employment_type? → Position=:Position, Placement =: Placement, Title =: Title, Employment_type =: Employment_type

But not all of these equations define most of the rules in the Title. The emergence of this multi-entity can be a combination of JSP. Another complex of Requirements is rule-based with a combination of hand-coded and learning-based. What we believe in this extraction requires bootstrapping-models [7] [8] to facilitate extraction. See the list in Table 5.

Statistical methods of entity extraction convert the extraction task to a problem of designing a decomposition of the unstructured text and then labeling various parts of the decomposition, either jointly or independently [6]. We believe that extracting Job Desk and Benefits as well as

Table 4.
Hand-coded Methodology

No.	Category	Job Standard Parameter
1	Company	Company
2.	Company	Industry
3.	Company	City
4.	Company	Phone
5.	Company	Website
6.	Company	Coordinate
7.	Jobs	Domain
8.	Jobs	Position
9.	Jobs	Placement
10.	Jobs	Employment Type
11.	Jobs Description	Education
12.	Jobs Description	Major
13.	Jobs Description	Experiences
14.	Jobs Description	Age
15.	Jobs Description	Gender
16.	Jobs Description	Marital Status

Table 5.
Job Standard Parameter for Jobs Descriptions

No.	Category	Job Standard Parameter
1	Address	Company
2.	Title	Industry
3.	Requirement	Education, Major, Experience, Soft-Skill, Hard-skill, Age, Gender, Marital Status

Multiple Entity in Requirements requires a complex methodology and ontology [7] for the extraction process. See the list on Table 5.

D. The Output of Extraction

We believe that the readers of this paper know the same field. So that each of them has a preference for how the extraction results are changed by the data type. However, this interesting thing lies in the relationship between Position and Experience. We identify some websites that display Experience with the same meaning as Position. Then, we also identified the strange behavior of JSP. We assume that the company has detailed candidate criteria. However, not all websites have the same features. Which finally forced all of the criteria to be placed in the Requirements.

IV. CONCLUSION

This study provides a detailed description of each component of information extraction on the job vacancy website in Indonesia. Starting from identifying the type of Structured Extraction to the output of extraction. This study also developed JSP from previous research. Each JSP is also explained in detail how this JSP should be extracted and how the challenges are. From this paper, data scientists can make pipeline more directed. In the future work, we will propose a hybrid web data extraction technique [8] that outperforms the limitations of supervised techniques in extracting job vacancy website in Indonesia.

REFERENCES

[1] G. V. A. N. Hoye and F. Lievens, "Investigating web-based recruitment sources: Employee testimonials vs word-of-mouse," *Int. J. Sel. Assess.*, vol. 15, no. 4, pp. 372–382, 2007, doi: https://doi.org/10.1111/j.1468-2389.2007.00396.x.

[2] P. Montuschi, V. Gatteschi, F. Lamberti, A. Sanna, and C. Demartini, "Job recruitment and job seeking processes: how technology can help," *IT Prof.*, vol. 16, no. 5, pp. 41–49, 2014, doi: 10.1109/MITP.2013.62.

[3] M. F. Koh and Y. C. Chew, "Intelligent job matching with self-learning recommendation engine," *Procedia Manuf.*, vol. 3, pp. 1959–1965, 2015, doi: 10.1016/j.promfg.2015.07.241.

[4] K. Adnan and R. Akbar, "Limitations of information extraction methods and techniques for heterogeneous unstructured big data," *Int. J. Eng. Bus. Manag.*, vol. 11, pp. 1–23, 2019, doi: 10.1177/1847979019890771.

[5] D. Celik *et al.*, "Towards an Information Extraction System Based on Ontology to Match Resumes and Jobs," in *Proceedings - International Computer Software and Applications Conference*, 2013, pp. 333–338, doi: 10.1109/COMPSACW.2013.60.

[6] S. Sarawagi, "Information extraction," *trends r databases*, vol. 1, no. 3, pp. 261–377, 2007, doi: 10.1561/1500000003.

[7] J. Zhang and W. Z. Ding, "An Improved Ontology-Based Web Information Extraction," in *Proceedings - 2015 International Conference of Educational Innovation Through Technology, EITT 2015*, 2016, pp. 37–41, doi: 10.1109/EITT.2015.14.

[8] F. Gutierrez, D. Dou, S. Fickas, D. Wimalasuriya, and H. Zong, "A hybrid ontology-based information extraction system," *J. Intell. Mater. Syst. Struct.*, vol. 42, no. 6, pp. 798–820, 2016, doi: 10.1177/1045389X14554132.