Correlation Measurement Between UNSPSC and KBLI 2009 Based on Classification

Edi Wahyu Widodo[#1], Tri Harsono[*2] and Ali Ridho Barakbah[*3]

[#] *Study Program of Applied Master's Degree, Information and Computer Engineering, Politeknik Elektronika Negeri Surabaya, Surabaya, 60111, Indonesia*
** Politeknik Elektronika Negeri Surabaya, Surabaya, 60111, Indonesia*

Abstract *— Electronic world have penetrated almost all fields, including in the field of procurement of goods and services by the government in Indonesia. Since 2007 in Indonesia has implemented the electronic procurement (e-procurement) of goods and services. In the procurement of goods and services that are now running, there is no classification or standard of goods and services as well as existing enterprises. In order to become more professional in auction, it is necessary to implement the classification. Classification of goods and services in accordance with the UNSPSC and business classification in Indonesia using KBLI edition 2009 to find which the business classification according to the UNSPSC to be in the auction, used methods of correlation measurements with text mining, allowing to find appropriate business classifications.*

Keywords *— UNSPSC, KBLI 2009, E-Tendering, Text Mining, Correlation Measurements.*

I. INTRODUCTION

Procurement of goods and services is a routine activity that is always carried out by an agency, in this case the government agency. It aims to complement the needs of that agency or perhaps for the rejuvenation of the existing tools or equipment.

Existing procurement processes still neatly arranged or implemented. Therefore, to achieve effective procurement and efficient then be made to a system of procurement of goods and services that are better, which uses electronic media and also that the procurement process can be run properly. The existing system has been adopted from the system that has been and is running on the Surabaya city government. In Indonesia, government e-procurement of goods and services has been conducted since 2007 with a system embraced from the electronic procurement of goods and services that is owned by the city government of Surabaya.

Procurement system adopted is still no complete. Where in fact they merely change from a manual to an electronic digital or be with the aim of realization of the disclosure of information for all those who want to know information about the procurement of goods and services or this auction.

LKPP or Procurement Agency is the agency responsible for handling the procurement of government goods and services nationwide. LKPP are responsible for both regulation and the running system. Until 2015, the system running quite well and have been used more than 600 government agencies throughout Indonesia.

On this system, there has been no classification of types of goods and services as well as classification of the type of business fields in Indonesia. This causes every field of business can follow all kinds of goods and services held by the government and did not result in a business field focuses on a field. A vendor that does not focus on the type of business that was involved field makes the vendor has a field that is too broad to be less professional or expert in one field there.

II. PREVIOUS WORK

SPSE or an electronic procurement system is a system used to perform the procurement or sale of goods or services by the Indonesian government. This system was first developed by the city government of Surabaya and later adopted by the Indonesian government since 2007 until now. This system has been used on more than 600 government agencies throughout Indonesia under the supervision LKPP as an institution responsible for the procurement of government goods and services.

Figure 1. Current E-Procurement System[1]

So far the system has evolved to version 3.6. Until this version, new auctions generally be classified into 4

types of auctions are: procurement, construction, consultancy and other services. So that the grouping has not been explicitly on a certain type of goods or services. As for the business classification in Indonesia based on KBLI 2009 have not been integrated in this system, only limited classification and have not implemented on the rule system.

With the running system now, the grouping of the goods and services are global and not specific. Causing multi perception of the goods or services that will be in the auction. It requires more time to give an explanation to prospective vendors that will follow the auction, so the auction will take more time to complete. Besides that, frequent inaccuracy of the specification of goods or services that have been in getting the results of the auction.

In addition, due to the unimplemented classification by existing businesses in Indonesia on this system, causing a vendor can participate in various types of auctions that exist. Which have resulted in the work imperfect because done by who are not experts and not his field. For example, there are companies registered as a seller of electronic goods but also can follow the procurement of goods by type of building materials. There are companies engaged in construction but can participate in the auction for the manufacture of consultancy services of computer software. Things like this should be eliminated in order to encourage the emergence of more professional vendor in the work.

III. System Overview

To overcome existing deficiencies in the previous system. We offer a solution in which there will be implemented the classification of goods and services based on UNSPSC (United Nations Standard Products and Services Code) and collaborate with Indonesian Standard Industrial Classification 2009 edition for later procurement of goods and services in Indonesia better.

A. UNSPSC

The United Nations Standard Products and Services Code® (UNSPSC®), managed by GS1 US™ for the UN Development Programme (UNDP), is an open, global, multi-sector standard for efficient, accurate classification of products and services. UNSPSC is an efficient, accurate and flexible classification system for achieving company-wide visibility of spend analysis, as well as, enabling procurement to deliver on cost-effectiveness demands and allowing full exploitation of electronic commerce capabilities. Encompassing a five level hierarchical classification codeset, UNSPSC enables expenditure analysis at grouping levels relevant to your needs. You can drill down or up to the codeset to see more or less detail as is necessary for business analysis[2].

The UNSPSC offers a single global classification system that can be used for:

- Company-wide visibility of spend analysis.
- Cost-effective procurement optimization.
- Full exploitation of electronic commerce capabilities.

How does UNSPSC work?

XX Segment

The logical aggregation of families for analytical purposes

XX Family

A commonly recognized group of inter-related commodity categories

XX Class

A group of commodities sharing common characteristics

XX Commodity

A group of substitutable products or services

XX Business Function

The function performed by an organization in support of the commodity.

All UNSPSC entities are further identified with an 8-digit structured numeric code which both indicates its location in the taxonomy and uniquely classifies it. An additional 2-digit suffix indicates the business function identifier. A structural view of the code set would look as follows:

Table 1. UNSPSC Example

Hierarchy	Category Number	Name
Segment	43	Information Technology Broadcasting and Telecommunications Communications Devices and Accessories
Family	20	Components for information technology or broadcasting or telecommunications Computer Equipment and Accessories
Class	15	Computers Computer accessories
Commodity	01	Computer switch boxes Docking stations
Business Function	14	Retail

B. KBLI 2009

Klasifikasi Baku Lapangan Usaha Indonesia or Indonesian Standard Industrial Classification (ISIC) is

a standard classification of economic activities are located in Indonesia. ISIC 2009 arranged to provide a set framework of a comprehensive classification of economic activities in Indonesia to be used for the uniform collection, processing, presentation and analysis of statistical data according to economic activity, as well as to study the economic situation and performance according to economic activities. With the uniformity, statistical data of economic activity can be compared to a standard format at the national, regional, and international levels[3].

Standard Industrial Classification of Indonesia Year 2009 was published in the form of Regulation of the Central Bureau of Statistics No. 57 Year 2009 on the Indonesian Standard Industrial Classification (BPS Perka No. 57 Year 2009 on KBLI) in December 2009. Perka BPS No. 57 Year 2009 is a revision of KBLI 2005. In its development based on input from various parties need to be improved and the addition of some business activities to the improvement of Perka BPS No. 57 year 2009.

KBLI 2009 classify the whole activity / economic activity in several fields of business activities which are distinguished based approach that emphasizes the process of economic activities in creating the goods / services, and approach more functions look at the functions of economic actors in creating the goods / services. The business unit is not differentiated according to the status of ownership, type of legal entity, or mode of operation. Production units which perform the same economic activity are classified in the same group of KBLI 2009, regardless of whether the production units are part of a legal entity or not, private or government, or individual, even if derived from enterprise consisting of more than the establishment or not.

KBLI 2009 Code Structure

a. **Category**, shows a line of staple classification of economic activities. This classification is coded one-digit code alphabet. In KBLI 2009, all economic activities in Indonesia are classified into 21 categories. The categories are coded letters from A to U.

b. **Principal Class**, a further description of the category. Each category is broken down into one or several base classes (up to five groups of goods, except for processing industry) according to the nature of each base class. Each subject was given a class of two-digit code.

c. **Class**, a further description of the principal class (point b). Class code consists of three digits, the first two digits indicate the principal class concerned, and the last digit indicates the economic activities of each class are concerned. Each base class can be broken down into as many as nine classes.

d. **Sub-class**, a further description of economic activity covered by a class (item c). Sub-class code consists of four digits, which is the first three-digit code shows the class-related, and the last digit indicates the economic activities of the Sub-class concerned. Each class can be broken down further into as many as nine sub-class.

e. **Groups**, is intended to break down the activities which are covered by a 'sub class' into several more homogeneous activities.

Table 2. KBLI 2009 Code Structure

KBLI 2009 Structure		Qty
Category	(alphabet)	21
Principal Class	(2 digits)	88
Class	(3 digits)	241
Sub class	(4 digits)	514
Groups	(5 digits)	1457

C. Text Mining

Just as data mining can be loosely described as looking for patterns in data, text mining is about looking for patterns in text[4]. However, the superficial similarity between the two conceals real differences. Data mining can be more fully characterized as the extraction of implicit, previously unknown, and potentially useful information from data. The information is implicit in the input data: it is hidden, unknown, and could hardly be extracted without recourse to automatic techniques of data mining. With text mining, however, the information to be extracted is clearly and explicitly stated in the text. It's not hidden at all—most authors go to great pains to make sure that they express themselves clearly and unambiguously and, from a human point of view, the only sense in which it is "previously unknown" is that human resource restrictions make it infeasible for people to read the text themselves. The problem, of course, is that the information is not couched in a manner that is amenable to automatic processing. Text mining strives to bring it out of the text in a form that is suitable for consumption by computers directly, with no need for a human intermediary[5].

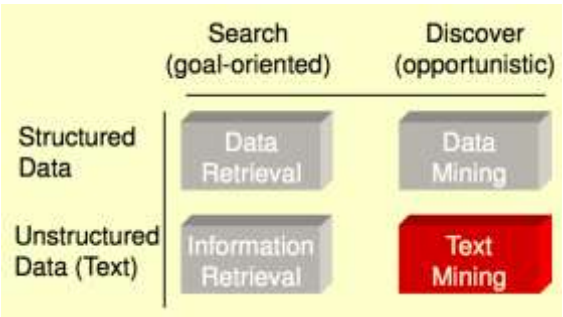

Figure 2. Text Mining

The text mining process have a few part, such as tokenizing, filtering, stemming/lemmatization, tagging, analyzing.

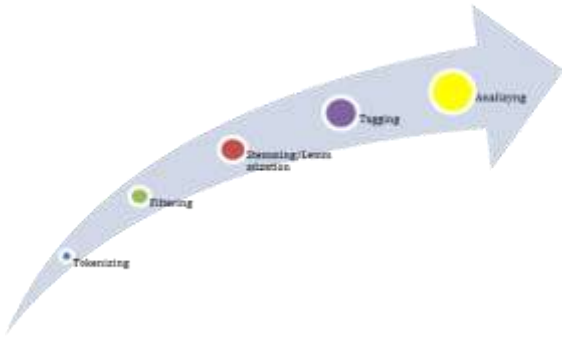

Figure 3. Text Mining Process

Tokenizing

Tokenization, or splitting the input text into words, is an important first step that seems very easy but is fraught with small decisions: how to deal with apostrophes and hyphens, capitalization, punctuation, numbers, alphanumeric strings, whether the amount of white space is significant, whether to impose a maximum length on tokens, what to do with non-printing characters, and so on. It may be beneficial to perform some rudimentary morphological analysis on the tokens[6].

Tokenization is a critical activity in any information retrieval model, which is simply segregates all the words, numbers, and their characters etc. from given text or document and these identified words, numbers, and other characters are called tokens.

Filtering

Filtering helps to provide the flexibility when we want to design data sources and mining structure so that a single mining structure can be created based on the comprehensive data source view. For training and testing different models, filters can be created to use only a part of that data and no need to build a different structure for each subset of data. We can use filter by length, Content, Indonesian, dictionary and Region etc. Filtering stage is the stage of taking important words from the tokens we have created. Could use the algorithm stop list (discard the less important word) or word list (save important words)[7].

In this paper tokens are filtered by length and words stop list. This operator filters tokens based on their length (i.e. the number of characters they contain) and stop list words. Parameters used in this operator to check length are:

- min chars:- The minimal number of characters that a token must contain to be considered.
- max chars:- The maximal number of characters that a token must contain to be considered.

Stemming

Stemming also known as lemmatization is a technique for the reduction of words into their stems, base or root. Many words in the Indonesian language can be reduced to their base form or stem e.g. Moreover, names can be transformed into root by removing the "s", for e.g., During the stemming process the variation "Stem's" in a sentence is reduced to "Stem" and this removal may lead to an incorrect stem or root. However, if the words are not used for human interaction then, these stems do not have to be a problem for the stemming process. But the stem is still useful, because all other inflections of the root are transformed into the same root.

In linguistic morphology and information retrieval, stemming is the process for reducing inflected (or sometimes derived) words to their stem, base or root form—generally a written word form. The stem need not be identical to the morphological root of the word; it is usually sufficient that related words map to the same stem, even if this stem is not in itself a valid root. Algorithms for stemming have been studied in computer science since the 1960s. Many search engines treat words with the same stem as synonyms as a kind of query expansion, a process called conflation.

Lemmatisation (or lemmatization) in linguistics, is the process of grouping together the different inflected forms of a word so they can be analysed as a single item.

In computational linguistics, lemmatisation is the algorithmic process of determining the lemma for a given word. Since the process may involve complex tasks such as understanding context and determining the part of speech of a word in a sentence (requiring, for example, knowledge of the grammar of a language) it can be a hard task to implement a lemmatiser for a new language.

In many languages, words appear in several inflected forms. For example, in English, the verb 'to walk' may appear as 'walk', 'walked', 'walks', 'walking'. The base form, 'walk', that one might look up in a dictionary, is called the lemma for the word. The combination of the base form with the part of speech is often called the lexeme of the word.

Lemmatisation is closely related to stemming. The difference is that a stemmer operates on a single word without knowledge of the context, and therefore cannot discriminate between words which have different meanings depending on part of speech. However, stemmers are typically easier to implement and run faster, and the reduced accuracy may not matter for some applications[8].

Tagging

Tagging stage is the stage of seeking an early form / root of each word or words past results stemming. Some words are usually undergo morphological changes shape into another word. for that there is the need to return it to its original shape

Analyzing

Analyzing stage is the stage of determining how much connectivity between words between documents or text which mined. Term Frequency-Inversed Document Frequency (IDF TF-) is the most simple algorithms that are usually used for scoring[9]. In this text mining, we have developed a new method/formula to give a score to each document. The formula is described below

$$Vfkmax_n = \frac{SW_n}{k}$$

If $fk_{n,i} < Vfk_{max}$

$$r_{n,i} = \frac{fk_{n,i}}{SW_n}$$

else

$$r_{n,i} = \frac{Vfkmax_n}{SW_n} + \frac{fk_{n,i} - Vfkmax_n}{SW_n \,.k}$$

$$R_n = \sum_{i=1}^{n} r_{n,i}$$

$r_{n,i}$ = score of each keyword i in document n
n = total number of documents
k = number of many keywords
$fk_{n,i}$ = number of found keyword i in document n
SW_n = number of words in document n
$Vfkmax_n$ = Maximum value of each keyword in document n
R_n = Score of document n

The formula above has been tried and compared to some other method for example such as TF-IDF and TF.

IV. Experiment Result

System is tested on another system with combined way. This system is a small part of a larger system that is e-procurement which currently still in the design stage.

Experiments carried out first by entering keywords to search for goods or services that exist in the list of UNSPSC and after that it will appear UNSPSC which has the closest relationship to the keywords in question.

Having obtained the appropriate UNSPSC code, it will look for a suitable business classification to UNSPSC codes that have been selected by using correlation measurements.

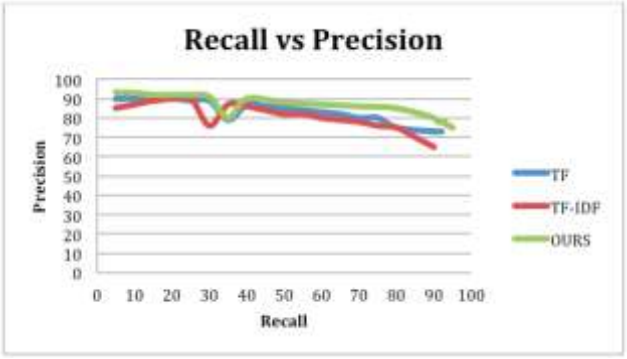

Figure 4. Text Mining Recall vs Precision

Based on some of the results of experiments that have been done, we developed a new method is more suitable for this case in comparison with other methods.

References

[1] LKPP Indonesia
http://lpse.lkpp.go.id

[2] UNSPSC
http://www.unspsc.org

[3] Head of The Central Statistics Regulation Number 57 of 2009.

[4] Munindar P. Singh, "The Practical Handbook of Internet Computing", CHAPMAN & HALL/CRC Computer and Information Science Series, 2005.

[5] Ian H. Witten, "Text Mining", Computer Science, University of Waikato, Hamilton, New Zealand, 2005.

[6] Vikram Singh and Balwinder Saini, "An Effective Tokenization Algorithm For Information Retrieval Systems", Departement of Computer Engineering, National Institute of Technology Kurukshetra, Haryana, India, 2014.

[7] Tanu Verma and Renu and Deepti Gaur, "Tokenization and Filtering Process in RapidMiner", International Journal of Applied Information System (IJAIS), 2014.

[8] Text Mining Online
http://www.textminingonline.com

[9] Joel Larocca Neto and Alexandre D. Santos and Celso A.A. Kaestner and Neto Alexandre and D. Santos and Celso A. A and Kaestner Alex and Alex A. Freitas, "Document Clustering and Text Summarization", Pontificia Universidade Catolica do Parana, Postgraduate Program in Applied Computer Science, Brazil, 2000.

[10] Martin Hepp, Joerg Leukel, and Volker Schmitz, "A Quantitative Analysis of Product Categorization Standards: Content, Coverage, and Maintenance of eCl@ss, UNSPSC, eOTD, and the RosettaNet Technical Dictionary", Springer, 2006

[11] Leo Obrst Robert E. Wray Howard Liu, "Leo Obrst Robert E. Wray Howard LiuOntological Engineering for B2B E-Commerce", Proceedings of the international conference on Formal Ontology in Information Systems, 2001

[12] Bjornar Larsen and Chinatsu Aone, "Fast and effective text mining using linear-time document clustering", Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining, 1999

[13] Weiguo Fan, Linda Wallace, Stephanie Rich, Zhongju Zhang, "Tapping the power of text mining", Magazine Communications of the ACM - Privacy and security in highly dynamic systems, Volume 49 Issue 9, September 2006

[14] Tuomo Kakkonen and Tabish Mufti, "Developing and applying a company, product and business event ontology for text mining", Proceedings of the 11th International

Conference on Knowledge Management and Knowledge Technologies, 2011

[15] Yue Dai, Tuomo Kakkonen, Erkki Sutinen, "MinEDec: a Decision-Support Model That Combines Text-Mining Technologies with Two Competitive Intelligence Analysis Methods", International Journal of Computer Information Systems and Industrial Management Applications, 2011

[16] Kalesha, Pattan, M. Babu Rao, and Ch Kavitha. "Efficient Preprocessing and Patterns Identification Approach for Text Mining." INTERNATIONAL JOURNAL OF COMPUTER TRENDS & TECHNOLOGY 1.6: 124-129.

[17] Saritha, A., and N. NaveenKumar. "Effective Classification of Text." International Journal of Computer Trends *0* 20: 40-60.

[18] Ababneh, Jafar, et al. "Vector space models to classify Arabic text." International Journal of Computer Trends and Technology (IJCTT) 7.4 (2014): 219-223.