Indonesian Journal of Electrical Engineering and Computer Science
Vol. 6, No. 1, April 2017, pp. 50 ~ 57
DOI: 10.11591/ijeecs.v6.i1.pp50-57

Mining Relation Extraction Based on Pattern Learning Approach

Mujiono Sadikin
Faculty of Computer Science, University of Mercu Buana, Jl. Meruya Selatan No. 1, Kembangan, Kebon Jeruk, Jakarta, Indonesia
Corresponding author, e-mail: mujiono.sadikin@mercubuana.ac.id

Abstract

Semantically, objects in unstructured document are related each other to perform a certain entity relation. This certain entity relation such: drug-drug interaction through their compounds, buyer-seller relationship through the goods or services, etc. Motivated by that kind of interaction, this study proposes a method to extract those objects and their interactions. It is presented a general framework of object-interaction mining of large corpora. The framework is started with the initial step in extracting a single object in the unstructured document. In this study, the initial step is a pattern learning method that is applied to drug-label documents to extract drug-names. We utilize an existing external knowledge to identify a certain regular expressions surrounding the targeted object and the probabilities of that regular expression, to perform the pattern learning process. The performance of this pattern learning approach is promising to apply in this relation extraction area. As presented in the results of this study, the best f-score performance of this method is 0.78 f-score. With adjusting of some parameters and or improving the method, the performance can be potentially improved

Keywords*: object interaction, object relation, information extraction, pattern learning, drug-name*

1. Introduction

A contained object interaction is one of the knowledge that can be extracted from a corpus collection. Several kinds of interactions can be found, for example interactions among drugs, interactions among family members, buyer-supplier interactions or interaction between employer and job seeker. Studies concerning interactions have been performed by many researchers. Drug–drug interactions mined from unstructured or structured documents was published in some papers such as: [1–9] The study of different object interactions contained in digital library document is presented in, whereas in [10], author studied the interaction between researchers or conference participants.

The study of drug–drug interaction (DDI) as proposed by Takarabe, et al [1] is approached by classification-based method, an ATC (Anatomical Classification Chemical). Whereas in the second study, Takarabe et al, the analysis of DDI networks are performed through the extraction of information contained in drug label such as drug-effects, indication, contraindication and enzyme ingredients. Both of these studies use KEGG-BRITE (a structured database) to fulfil their purpose. The other classification-based method is proposed by Mujiono et al. that use the DOEN 2011 as the base for classification method and FDA-Drugs as the data sources. Utilization of machine learning approach to predict DDI is proposed by He Z et al [11] that also used KEGG-BRITE database. NLP is the most wide used approache in text mining [12]. As used in the othe area, NLP technique is also applied in drug related extraction such as proposed by Jacinto Mata et al [4]. In the studi authors combine machine learning and NLP extract DDI information contained in a DDI corpus. Several other methods and data sources related to DDI studies are: Interaction Profile Fingerprint applied to DrugBank database by Vilar S et al [5], ANN applied to KEGG BRITE, BRENDA, SuperTarget, and DrugBank by Polak et al [7].

Most of the proposed methods are applied to structured data sources, i.e. database, while at the same time, there are many unstructured data source that contains the object interaction information. The application of the classification-based method has a limitation, since it can be applied only to object entities which have clearly defined their class. On the other

Received December 21, 2016; Revised March 1, 2017; Accepted March 15, 2017

hand, the limitation of the NLP-based method is its high dependency to a specific human language. NLP based also requires a knowledge resources as the central repository which is not always available in any case [13].

The study presents the authors' proposed general framework to extract network interactions between object entities contained in unstructured data, i.e. corpora collection. The proposed method is also independent to any specific natural language and as well as any classification-based approach. In the initial stage of the study, presented in this paper, the authors show the results and analysis of the initial experiment of pattern learning method that is applied to drug-label to extract drug-name.

2. Related Study

2.1. Object Interaction

One of the topics surrounding interaction among objects is drug – drug interaction. Several studies regarding DDI are described in this section. Jacinto, M. [4] propose machine learning approach such: SVM, Naive Bayes, Decision Tree, and Adaboost to extract DDI contained in English drug label. The Chi-square feature selection mechanism is applied to handle those document's high dimensionality. The detection of drug – protein interaction by utilizing a bipartite-graph learning-method was proposed by Yamanashi Y.et al [6]. This method treats the drugs, its protein target and interaction between them as a pharmacological space which is represented as a bi-party graph. The new predicted drug – target is estimated by calculating their closeness in the space, based on a predefined threshold distance. Yamanashi used structure drug data base such as KEGG BRITE, BRENDA, SuperTarget, and DrugBank, to do the experiments. Polak et al [7] applied ANN approach to predict DDI by using drug-interaction information and their chemical compound dataset.

Several other research concerning object-interaction mining are object interactions contained in digital library [10], interaction between researchers or conference participants based on the content of on-line media [14]. In the first publication, Yizhou Sun et al proposed a GLM (Generalized Linear Model) based supervised framework to perform a relationship building time model. The objects in this study were: author, venue, term, and people that are extracted from DBLP.

2.2. Rule Based Object Extraction

Information or objects extraction of unstructured text document is one of the most challenging study in the text mining area. Due to the increasing of corpora, the growing of human natural language, and the unstructured formatted data, the difficulties of information extraction is increasing in the future [15]. Object extraction has been studied in many different contexts and purposes. Several methods, rule based learning or statistical based method have been published [15]. Some of the rule based learning are presented in this section, as this study also uses this approach. One of the rule based learning methods is bootstrapping approach which has been studied by many researchers. The first generation of bootstrapping approach is Snowball [16]. In this study, the author applied bootstrapping technique for extraction of binary relations, such as Organization-Location, e.g., between Microsoft and Redmond, WA. Thellen M and Riloff E [14] proposed bootstrapping method to infer semantic lexicon of new words. The bootstrapping method was utilized to perform new patterns to identify the new word category. Pattern learning with bootstrapping approach is also studied by W. Lin et al [17] to extract the names entities of a certain domain. In this case, the author applied the pattern learning algorithm for disease and location category. Another pattern learning approach to extract information of specific domain was proposed in [18].

One of the most recent study related to bootstrapping is proposed by Liu, Ting and Tomek S [19]. In their study, the bootstrapping method extracted events and its relation from text, based on the pattern resulted from learning process. These learning processes include two mechanisms: learning through pattern mutation and learning by exploiting structural duality. Event information viewed as multiple face, which extracted from news is also studied with bootstrapping approach. This kind of learning pattern – bootstrapping application was published in [20].

3. Approaching Method

3.1. Aim

The goal of the overall study is to predict the potential interaction between entities extracted from unstructured text. To achieve this goal, the approach defines a relation between the main object and supporting object. In this context the main objects, for example. are seller, buyer, drug, etc. Whereas the supporting objects are goods or services that were sold by the seller or bought by the buyer, drug chemistry compound if the main object is drug or children if the main object is father or mother.

3.2. General Frameworks

The author proposed general framework consists of two main phases. The first phase is the stage to extract main object-supporting object relation, and the second phase is to group those relations extracted by categorical bi-clustering approach as illustrated in Figure 1. To extract object relation in stage one, the authors adopt and enhance bootstrapping methods as proposed in [21], [22], and [23]. The enhancements of the authors' proposal are in the pattern generation methods and the use of independent format of external knowledge. More detail regarding to the first stage and its initial experiments is presented in this paper.

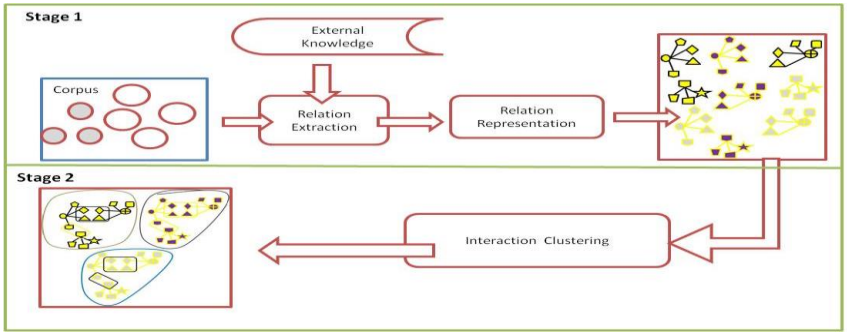

Figure 1. The Study Global Framework

3.3. Pattern Learning for Object Extraction

In the first experiment of the first stage the authors apply the object extraction approach to extract drug name from drug label document. The pattern learning approach proposed in this study uses the Indonesian WordNet published by the PAN Localization project [24] as an external knowledge. The Indonesian WordNet is a collection of more than 1.000.000 words in Bahasa Indonesia, which collected from various sources, such as news agencies, on-line media publishers, Internet blogs, websites and others [24]. Based on the assumption that the name of drugs distributed in Indonesia are mostly unique and are not commonly used in daily term in Bahasa Indonesia, the authors use the Indonesian Wordnet as a guide to determine if such certain word identified in data set is a drug name or not. The drug name object extraction framework is illustrated in Figure 2. Another input of this framework, in addition of the WordNet, is the initial pattern constructed manually.

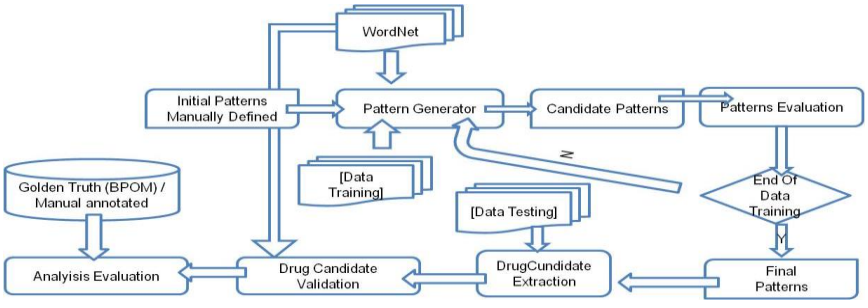

Figure 2. Object extraction by Pattern Learning Framework

3.4. Pattern Generation

To extract an object relation, the authors define an object relation as illustrated in Figure 3. A relation includes one main object and one or more its supporting object. In certain document the relation between the main object and its supporting objects is marked by relation term. The relation term is a certain word, commonly a verb, which describes what kind of relation between the main object and its supporting objects. For example, in a drug label document drug–drug component relation, drug is a main object, its chemical components are the supporting object and "komposisi" (composition) is relation term.

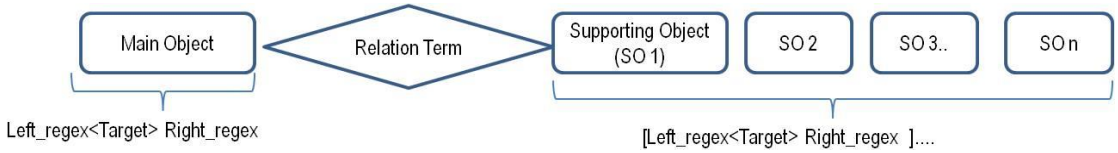

Figure 3. Object Relation Representation

Both of the main objects and supporting objects is identified by a target which are surrounded by a kind of regular expression (regex). The regex consists of left regex and right regex. The pattern is the combination of those regexes. In this approach, pattern generation is performed by identifying certain words which are not contained in the WordNet. Once the words are identified, than the other words located on its left and right are used as a regular expression of the candidate pattern.

3.5. Pattern Selection

The number of candidates' pattern provided in the training phase can be very high. The maximum number of these candidates pattern is as many as the number of training document, if each training document contains a unique pattern. If we use those entire patterns resulted from a large number of documents, it will consume too many resources (time, processor, or storage). To reduce the size of resources requested, we select only candidate pattern that fulfils a certain score. This score pattern is defined based on the probabilities of left and right regular expression surround the target. This next paragraph describes more formal definition regarding to the scoring method. We executed two scoring techniques.

Pattern Scoring #1

Definition

Each pattern consists of Left Tuple (LT), Relation-Term, and Right Tuple (RT)
LT-L = $\{(lt\text{-}l)_1, (lt\text{-}l)_2, \ldots (lt\text{-}l)_n\}$; $(lt\text{-}l)_j$ is j^{th} left regex of the left-tuple
N = Quantity of LT-L, N >= n ; N = n if $(flt\text{-}l)_j$ = 1 for $0<j\leq n$
$(flt\text{-}l)_j$ = frequency of $(lt\text{-}l)_j$ in the training set, then

$$P_j(lt\text{-}l) = (flt\text{-}l)_j / N, \quad (1)$$

LT-R = $\{(lt\text{-}r)_1, (lt\text{-}r)_2, \ldots (lt\text{-}r)_m\}$; $(lt\text{-}r)_i$ is k^{th} right regex of the left-tuple
M = Quantity of LT-R, M >= m; M = m if $(flt\text{-}r)_k$ = 1 for $0<k \leq m$
$(flt\text{-}r)_k$ = frequency of $(lt\text{-}r)_k$ in training set, then

$$P_k(lt\text{-}r) = (flt\text{-}r)k / M, \quad (2)$$

Pattern Scoring (PS) (Tuple Probabilities = Join probabilities of left regex and right regex. For the left tuple, the pattern score is:

$$PS_i = P_i\ (Left\text{-}Tuple_i) = P_j(lt\text{-}l) * P_k(lt\text{-}r)\ ; \quad (3)$$

$\forall i,\ i \in \{1,2...n*m\}$; $\forall j,\ j \in \{1,2...n\}$; $\forall k,\ k \in \{1,2...m\}$;

Pattern Scoring #2

The next pattern scoring method used in this experiment is performed by adjusting equation (3). By treating the left regex and the right regex as independent variable to each other as the equation above, the number of the left pattern that can be generated are = $J * K$, with J is the number of the left regex of the left tuple and K is the number of the right regex of the left tuple. In the second pattern scoring method, not all of the right regex are paired with each left regex, but some of the right regexes are belong to certain left regex. This formulation is similar to conditional probabilities formulation. The formal definition of the second pattern scoring method, PS_i, is:

$$PS_i = Pi((lt\text{-}l)_j|(lt\text{-}r)_k), \quad (4)$$

$\exists i,\ i \in \{1,2...n*m\},\ i \leq s \leq n*m;\ \forall j,\ j \in \{1,2...n\}\ ;\ \exists k,\ k \in \{1,2...m\}, k \leq m;$

Algorithm

Based on the object extraction framework and the pattern scoring, the algorithm to generate new pattern using the training data set is presented in this section. The same skeleton of the algorithm is used both for the first or second pattern scoring, its difference is only in the update score mechanism block. This presented algorithm is based on the second pattern scoring.

```
Input (training_set, WordNet, initialPattern)
Output (NewPatternList <Left_Regex, Target, Right_Regex, Prob. of Pattern>)
Algorithm
  NewPatternList ←Null
  for all documents in training_set do
    if the document contains initialPattern.relation-term
      then
      Get-term-in left of relation-term that is
        not in WordNet
      Get Left_Regex, count its frequency
      Get Right_Regex, count its frequency for
        the Left_Regex
    for all pairs of Left_Regex and Right_Regex do
      Perform NewPatternList and Calculate P(R|L)
  Reorder on P(R|L) New Pattern in NewPatternList
  Output NewPatternList
```

4. Materials and Evaluations

4.1. Data Set and Pre Processing

To validate the pattern learning approach proposed, it is used drugs label documents that grabbed from various drug producers and regulator Internet sites located in Indonesia. Those sites are: http://www.kalbemed.com/, http://www.dechacare.com/, http://infoobatindonesia.com/obat/, and http://www.pom.go.id/webreg/index.php/home/produk/01. The drug labels are written in Bahasa Indonesia and their common content are drug name, drug components, indication, contra indication, dosage, and warning. Since almost all those documents are grabbed by the engine, their format is in htm or html. To filter the real content which contains information regarding drug label, we use html parser provided by http://sourceforge.net/projects/htmlparser/. The ground truth of the data test is performed manually. Drug name and drug component of the ground truth are annotated by expert.

4.2. Evaluation

To evaluate the performance of the object extraction pattern learning method performance, we use the common criteria in data mining: precision, recall, and f-score. Let $C = \{C_1, C_2, C_3...Cn\}$ is a set of drug-name extracted by this method against drug-label document set

D, and *K*= *{K1, K2, K3,...Kl}* is set of actual drug-name in document set *D*. Adapted from [25], those three criteria computed as follows:

$$\text{Precision}(Ki, Cj) = \frac{\text{True Positive}}{\text{True Positive+False Positive}} = \frac{|\text{Ki} \cap \text{Cj}|}{|\text{Cj}|} \quad (5)$$

$$\text{Recall}(Ki, Cj) = \frac{\text{True Positive}}{\text{True Positive+False Negative}} = \frac{|\text{Ki} \cap \text{Cj}|}{|\text{Ki}|} \quad (6)$$

where |*Ki*|, |*Cj*|, and |*Ki* ∩ *Cj*| denote the number of drug-name in K, in C, and in both K and C respectively. The computation of F-Score is performed by the formula:

$$\text{F} - \text{Score}\ (Ki, Cj) = \frac{2 * \text{Precision (Ki,Cj)} * \text{Recall (Ki,Cj)}}{\text{Precision (Ki,Cj)} + \text{Recall (Ki,Cj)}} \quad (7)$$

5. Experiments, Results and Discussion

5.1. Experiment Scenario

The experiment scenario is arranged based on: 1. pattern scoring technique, and 2. the volume of data set. The execution of both of that pattern scoring technique generates many patterns candidates. The candidate patterns generated are sorted in a descending order on its pattern score. Intuitively, it can be seen that the patterns scoring technique #1 generate more patterns than the pattern scoring technique #2. To evaluate the performance of the patterns scoring technique, we take the N top of generated patterns. N is the total number of patterns generated by pattern scoring technique #2. Then the better result, in this case is the pattern scoring technique #2, is used for the second experiments by adjusting the data set volume parameter. For both of the scenario, the data set is split into two parts, one part is as training set and the other as a test set by K-fold cross validation with K = 10 respectively. So there are 10 iterations for each scenario.

5.2. Results & Discussion

5.2.1 Pattern Scoring Technique Scenario

The performance of both pattern scenario techniques is illustrated in figure 4. In overall, the second technique is better than the first one. The minimum, maximum, and average of first technique f-score are: 0.204225, 0.338862, and 0.269687 whereas the second technique is: 0.581335, 0.444444, and 0.709677. The result is intuitive, since not all of the patterns generated by first technique is used for drug-name target extraction. In this experiment the average of the total number of patterns generated by the Pattern Scoring technique #1 is 101 whereas that is generated by Pattern Scoring technique #2 is 498.

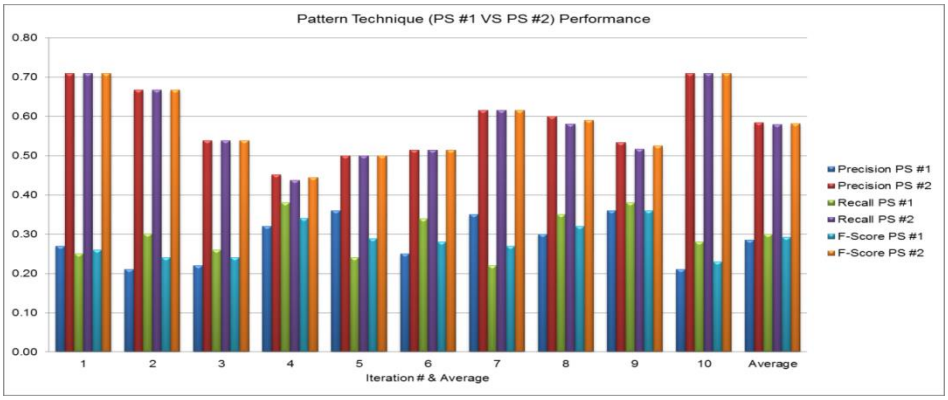

Figure 4. Performance of pattern scoring technique

5.2.2. Data Set Volume Adjustment Scenario

In the second scenario we compare the performance of the pattern generated by adjusting the quantity of data set which the first experiment uses to 340 drugs-label documents and the second uses to 900 drugs-labels. As a result of this experiment, it is shown that the more data set used, its performance is better on average. Figure 5 presents both of the performances. The performances of 900-drug labels are better in precision, recall, and f-score from 7 of 10 iterations.

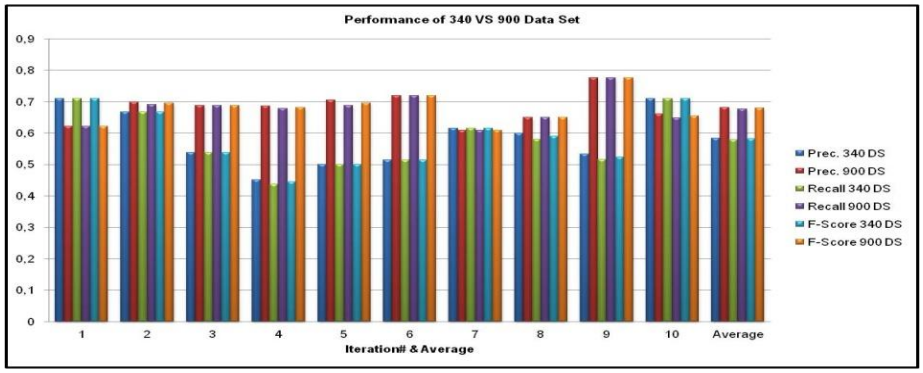

Figure 5. Performance of the 340 Data set vs 900 Data set

6. Conclusion and Future Study

This study presents a global framework to predict the potential information or object interaction contained in unstructured documents with an initial step to extract a certain object, based on pattern learning. The initial experiment shows that bootstrapped pattern learning approach still has open opportunity that can be exploited. Various pattern scoring techniques as a base of pattern selection that applied to data test need to be explored further. In this study the performance of patterns learning method is influenced by the volume of data set and as well as the technique, as indicated by the experimental results.

As an initial step, this study still opens many challenges to be exploited. There will be many tasks that can be explored vertically on the method and technique or horizontally on its application. In vertical point of view, the author will explore some task, such as extending the target not only to the main object, which is the drug-name, but also to its supporting object, which is the drug-component. The authors will also explore to extend the number of words surrounding the target from 1 to 1+n to get better performance. The recursion mechanism in obtaining the newer pattern list is also one of the many challenges. By the recursion, the pattern generated is used as an initial pattern to the next step. The other external knowledge, such as drug data bank such as RxNorm [26], NCI [27], or FDA [28] can also be assessed to be used as a pattern generation guidance. In the next step, the authors will explore the possibility of using internal knowledge come from the data set rather than external knowledge as applied in this study. At the horizontal point of view, the authors plan to apply the framework to others domain, such as trading area, so the sellers and buyers contained in the corpora can be extracted.

References

[1] Takarabe M, Okuda S, Itoh M, et al. Network analysis of adverse drug interactions. *Genome Inform* 2008; 20: 252–9.

[2] Takarabe M, Shigemizu D, Goto S, et al. Characterization and Classification Of Adverse Drug. *J Genome Inf.* 2010; 167–175.

[3] He Z, Zhang J, Shi X-H, et al. Predicting drug-target interaction networks based on functional groups and biological features. *PLoS One.* 2010; 5: e9603.

[4] Mata J, Santano R, Blanco D, et al. A Machine Learning Approach to Extract Drug-Drug Interactions in an Unbalanced Dataset. *1st Chall task Drug-Drug Interact Extr.* 2011: 6–12.

[5] Vilar S, Uriarte E, Santana L, et al. Detection of drug-drug interactions by modeling interaction profile fingerprints. *PLoS One.* 2013; 8: e58321.
[6] Yamanishi Y, Araki M, Gutteridge A, et al. Prediction of drug-target interaction networks from the integration of chemical and genomic spaces. *Bioinformatics.* 2008; 24: i232-40.
[7] Polak S, Brandys J, Mendyk a. *Neural System for in silico Drug-Drug Interaction Screening*. Int Conf Comput Intell Model Control Autom Int Conf Intell Agents, Web Technol Internet Commer. 2005; 2: 75-80.
[8] Sadikin M, Wasito I. *Translation and Classification Algorithm of FDA-Drugs to DOEN2011 Class Therapy to Estimate Drug-Drug Interaction*. In: The 2nd International Conference on Information Systems for Business Competitiveness 2013 (ICISBC 2013). Semarang Indonesia. 2013: 1–5.
[9] Boyce R, Gardner G. *Using Natural Language Processing to Identify Pharmacokinetic Drug-Drug Interactions Described in Drug Package Inserts*. In: the 2012 Workshop on Biomedical Natural Language Processing (BioNLP 2012). Montreal, Quebec, Canada. 2012: 206–213.
[10] Sun Y, Han J, Aggarwal CC, et al. When Will It Happen ? Relationship Prediction in Heterogeneous Information Networks. In: *WSDM'12*. Seattle, Washingtion, USA. 2012.
[11] He L, Yang Z, Lin H, et al. Drug name recognition in biomedical texts: A machine-learning-based method. *Drug Discov* Today. 2014; 19: 610–617.
[12] Naw N, Hlaing EE. Relevant Words Extraction Method for Recommendation System. *Bull Electr Eng Informatics.* 2013; 3: 680–685.
[13] Phyue Mandalay SU of CS. Unknown Word Detection via Syntax Analyzer. *IAES Int J Artif Intell*. 2http://iaesjournal.com/online/index.php/IJAI/article/view/1802 (2013).
[14] Xiang Zuo, Alvin Chin, Xiaoguang Fan, Bin Xu, Dezhi Hong, Ying Wang XW. *Connecting People at a Conference-A Study of Influence between Offline and Online Using a Mobile Social Application.* IEEE International Conference on Green Computing and Communications, Conference on Internet of Things, and Conference on Cyber, Physical and Social Computing. 2012: 277–284.
[15] Tang H, Ye J. *A Survey for Information Extraction Method*.
[16] Agichtein E, Gravano L. *Snowball: Extracting Relations from Large Plain-Text Collections*. In: DL '00 Proceedings of the fifth ACM conference on Digital libraries. New York, New York, USA, 2000: 85–94.
[17] Lin W, Yangarber R, Grishman R. *Bootstrapped Learning of Semantic Classes from Positive and Negative Examples*. Proceedings of the ICML-2003 Workshop on the Continuum from Labeled to Unlabeled Data. Washington DC. 2003.
[18] Patwardhan S, Riloff E. *Learning Domain-Specific Information Extraction Patterns from the Web*. In: Proceedings of the Workshop on Information Extraction beyond the Document. Sydney. 2006: 66–73.
[19] Liu T, Strzalkowski T. *Bootstrapping Events and Relations from Text Polish Academy of Sciences*. The 13th Conference of the European Chapter of the Association for Computational Linguistics. Avignon, France. 2012: 296-305.
[20] Huang R, Riloff E. *Multi-faceted Event Recognition with Bootstrapped Dictionaries*. In: Proceedings of NAACL-HLT 2013. Atlanta, Georgia. 2013: 41-51.
[21] Thelen M, Riloff E. *A bootstrapping method for learning semantic lexicons using extraction pattern contexts.* Proceedings of the ACL-02 conference on Empirical methods in natural language processing-EMNLP '02. Morristown, NJ, USA: Association for Computational Linguistics. 2002; 214-221.
[22] Sun A. *A Two-stage Bootstrapping Algorithm for Relation Extraction.* New York, NY, USA, 2009.
[23] Umamaheswari E, Geetha TV. *Learning Event Patterns from News Text Using Bootstrapping*. In: International Conference on Information System Security and Cognitive Science. Singapore, 2013: 48–54.
[24] Secretariat RP. PAN Localization Project, Indonesia country componentwww.pan10n.net (2010).
[25] Dagher GG, Fung BCM. Subject-based Semantic Document Clustering for Digital Forensic Investigations. *J Data Knowlege Eng*.1986.
[26] NLM. Unified Medical Language Systemhttp://www.nlm.nih.gov/research/umls/rxnorm/ (2014).
[27] NCI. NCI Drug Dictionaryhttp://www.cancer.gov/drugdictionary (2014).
[28] FDA. U.S. Food and Drug Administration. 2014; 1–6.