International Journal of Psychosocial Rehabilitation, Vol.24, Issue 01, 2020
ISSN: 1475-7192

Text Automation Classification Using the Lexical Feature Representation Method

[1] Murnawan, [2] R.A.E. Virgana, [3] Sri Lestari

Abstract— *The process of automating text classification plays an important role in organizing a text document, determining the characteristics and characteristics of a document. To determine a characteristic or information hidden in a large dataset is very necessary, this is because the unstructured document has many meanings, different meanings and purposes. Therefore, it is necessary to have a special method that can provide important information contained in a text document. The feature representation method that will be used in this research is N-Grams, as well as the use of bag of concepts which is the development of the concept of bag of words to reduce the level of computing in forming feature representations. The purpose of this research is to design a question categorization automation feature that does not have a category contained in a text document, by applying lexical feature representation concepts such as bag of concepts, bag of word and N-Gram to the question categorization automation feature. Based on the results of experiments on WEKA, the combination of lexical features between unigram, bigram, trigram and keyword from each category in the implementation of making data models using cross validation with a fold number of 10 shows that the combination of the bigram trigram and keyword features gives the percentage of instance properly classified more correctly high compared to other feature combinations that is equal to 96.5% with the J48 Tree classifier.*

Index Terms— *text classification, feature representation, bag of concept, n-gram*

I. INTRODUCTION

With the rapid development and growth of electronic information, a process is needed to solve a problem in order to make use of information that is meaningless at a glance into something that is very meaningful in the future.

One solution offered in dealing with these problems is by automating the process of text classification. Automation of text classification is very much needed in the digital era which is very fast nowadays because every day human life will always be dealing with a number of variants of text documents and the need for efforts to deal with large amounts of information which are often unstructured. The process of automating text classification plays an important role in organizing a text document, determining the characteristics and characteristics of a document.

To determine a characteristic or information hidden in a large dataset is very necessary, this is because the unstructured document has many meanings, different meanings and purposes. Therefore it is necessary to have a special method that can provide important information contained in a text document.

In the development of implementing a text classification automation, there are several examples of algorithms used such as Naive Bayes, Support Vector Machines or Decision Trees, and in this case there are also some lexical feature representations that influence the process of text classification automation to determine various information and characteristics of a document. texts such as N-Gram, traditional bag of word models, and also the bag of concept which is the development of the bag of word concept and various other feature representations.

[1] *Murnawan is with Information System Department, Widyatama University, Bandung, Indonesia (e-mail: murnawan@widyatama.ac.id)*
[2] *R.A.E. Virgana is with Information System Department, Widyatama University, Bandung, Indonesia (e-mail: rae.virgana@widyatama.ac.id)*
[3] *Sri Lestari is with Information System Department, Widyatama University, Bandung, Indonesia (e-mail: sri.lestari@widyatama.ac.id)*

The feature representation method that will be used in this research is N-Grams, as well as the use of bag of concepts which is the development of the concept of bag of words to reduce the level of computing in forming feature representations.

II. Literature Review

A. Artificial Intelligence

According to Russell and Norvig [1], AI or artificial intelligence is divided into several dimensional aspects which are represented as 4 windows, in the first aspect in line 1 relating to how the process of thinking and giving an opinion, then the second aspect in line 2 is related to behavioral behavior , in row 1 column focuses on humanism, then column 2 deals with rationality. The following representations are divided into 4 important aspects as following in figure 1.

Based on these representations it can be concluded that artificial intelligence is divided into 4 main categories namely, how a system can think like humans in general can understand a problem pattern and can provide solutions to problems, then how an existing system can think rationally based certain computations that provide valid and trusted results.

Thinking Humanly	**Thinking Rationally**
"The exciting new effort to make computers think ... machines with minds, in the full and literal sense." (Haugeland, 1985) "[The automation of] activities that we associate with human thinking, activities such as decision-making, problem solving, learning ..." (Bellman, 1978)	"The study of mental faculties through the use of computational models." (Charniak and McDermott, 1985) "The study of the computations that make it possible to perceive, reason, and act." (Winston, 1992)
Acting Humanly	**Acting Rationally**
"The art of creating machines that perform functions that require intelligence when performed by people." (Kurzweil, 1990) "The study of how to make computers do things at which, at the moment, people are better." (Rich and Knight, 1991)	"Computational Intelligence is the study of the design of intelligent agents." (Poole et al., 1998) "AI . . . is concerned with intelligent behavior in artifacts." (Nilsson, 1998)

Fig. 1. Some definitions of artificial intelligence, organized into four categories.

A system in artificial intelligence is required to be able to act like humans and must have several capabilities that meet the standards to create a system with good artificial intelligence, follows some examples of the implementation of the fields that become the capability obligations that must be met from an artificial intelligence system:

a. Natural Language Processing: through this field the system is expected to be able to communicate in any language used by humans.
b. Knowledge representation: the system is expected to be able to store various information that will be used in a core process in artificial intelligence.
c. Automated Reasoning: the system can use the information that has been stored to answer a given question or design a new conclusion.

d. Machine Learning: a system that has the ability to adapt to a new situation and create new patterns that are used to solve a problem.

B. Machine learning

Machine Learning is a study of algorithms to learn something in doing certain things done by humans automatically [2]. Learning in this regard is related to how to complete various existing tasks, or make a prediction of new conclusions that are accurate from various patterns that have been studied previously.

Machine Learning is one of the fields included in artificial intelligence that can affect various other aspects, namely static, mathematics and various theoretical aspects of computer science. Basically, Machine Learning has the goal to learn an algorithm to do a learning system automatically with a very minimal contribution made by humans in general.

Machine Learning is divided into 2 types of learning concepts [2][3]. First is supervised learning which is a machine learning technique that makes a function based on existing training data, in this case it can be said for this technique that is available in detail and well-classified training data that will be used a data model when testing is carried out with new test data and produces outputs that are as expected previously based on existing training data. Second is unsupervised learning which is a machine learning technique that seeks to represent an input pattern derived from training data and one of the differences with Supervised Learning is the lack of classification of input data. In machine learning the unsupervised learning technique becomes essential because the work system provided is the same as the workings of the human brain where in the learning process there are no role models or information and examples available to serve as models in conducting the testing process for solving a problem with new data.

Here are some examples of algorithms from the concept of Supervised Learning and Unsupervised Learning [4]:

1) Supervised Learning

a. Logistic Regression: is a statistical technique that is also used to solve problems by producing a prediction of new variables based on several variables that are already predetermined and looking for the relationship between input data variables that have been determined with output variables or new predictive variables.
b. K-Nearest Neighbors Algorithm (KNN): is a method used in solving the problem of classifying an object by implementing a feature space where an object that is training data and used as a data model is given a weighting value and represented in an n-dimensional vector. Then the problem solving is done by measuring the closest distance of the new object to the existing data model in the n-dimensional vector then the process of assigning a category to the new object.

2) Unsupervised Learning

Clustering: a method of forming the basic pattern of an object that is used to solve problems in the case of machine learning such as the classification or categorization of objects into a class / category. The formation of this basic pattern is done by using a number of features that are determined and this is done because the data used as a data model has not been defined by the data group so that it cannot determine the classification of an object, after getting the basic pattern formed then it can be used as a data model and used as training data in the case of new object categorization.

C. Natural language processing (NLP)

Natural Language Processing (NLP) is one branch of artificial intelligence that can translate between computers and human languages [5]. The method is used in computers to be able to understand in reading a line from a collection of words without providing a hint or study material for the computer in performing the translation calculation process.

NLP system has an input data in the form of a collection of several words into a sentence and will produce a structured representation in order to find a meaning from the input data given and the meaning generated from the process is used as output data.

There are several steps used in the NLP process [5]:

a. Morphological and Lexical Analysis
   It is a process that can carry out an analysis and identification of the morphological structure of a word, and is used to divide a text into a paragraph, collection of words, or a sentence.
b. Syntactic Analysis

It is a process of analysis of a collection of words in a sentence to identify the grammatical structure contained in the sentence. In this process a collection of words is carried out a transformation process into a structure that can describe how a collection of words interrelate with other words.

c. Semantic Analysis
Is a process that is used to carry out the final checking process in terms of checking the truth of the meaning that has been given to every word in the previous process, in this process the process of correcting relations and conformity with various other words.

d. Discourse Integration
The meaning of a sentence depends on the collection of words in the sentence, for example "He wants it" the word "that" is very dependent on the context of the sentence itself or other sentences.

e. Pragmatic Analysis
It is a process of separating or decreasing an existing sentence to produce an interpretation of an appropriate meaning, for example: "close the window?" should have interpreted the meaning of a request rather than a command.

D. Automatic text classification

Automatic Text Classification is one branch of artificial intelligence whose job is to carry out the process of grouping a text or document into a specific category or topic that has been determined.

Automatic Text Classification includes text classification based on topic and genre of text [6]. In general, the process of classifying a text is a process that groups a new text by using text that already has a category to be a comparison of the text to be grouped.

There is a process flow of general text classification as follows:

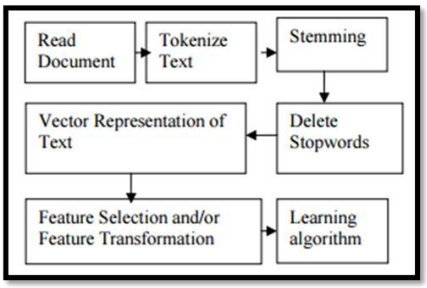

Fig. 2. Text Classification Process

There are 2 important parts of the process of classifying a text, namely preprocess and process data. In preprocess data there is a stemming process that is tasked to eliminate the affixes to the words in a text. Then there is the delete stopwords process whose task is to eliminate words that have no meaning at all in an existing document such as conjunctions (and, this, that, but etc.), or prepositions and the like.

Then in the Process Data section there is a Vector Representation of Text process which has the task of forming a feature which can be used to form a data model that will be used as a comparison material with new text that has not been given a category. Formation of a feature consists of various forms ranging from the lexical which does not pay attention to the sentence patterns of a text and semantics and syntactic that forms features based on the existence of certain relationships between a word with other words that can be used as a comparative feature in a document matrix that is formed between documents that will be used as a model with features that have been formed.

The next process is feature selection which is a process that helps to choose features that give a lot of influence in the classification process, it aims to reduce the high level of computation than before having features that do not provide many roles to be more concise.

In the next stage is the formation of a data model and the process of text classification using the Machine Learning algorithm to be tested with new data. The same process is carried out when receiving new data by forming a feature then a feature selection process is carried out and then in the next stage through the implementation of the Machine Learning algorithm there will be a calculation and comparison between the two data sets namely the data model with the new data which will then come out the results are classified documents.

E. Stemming algorithm

The Stemming algorithm used is the Nazief and Adriani stemming algorithm [7] which was coined by Nazief and Adriani who is a lecturer from the University of Indonesia who provided various new understanding and research on the stemming algorithm and was first described in a technical report that was not publicly opened on the Internet media. The workings of this algorithm are to use morphological rules to eliminate the prefix and suffix of a word and then use a base word database to be a comparison of whether it is the same between the words that the stem process wants to do and that is in the base word database.

The main focus of this algorithm is the use of basic word databases in Indonesian, the more complete the list of available words, the higher the accuracy given when using this algorithm.

Following are the steps in using the Nazief and Adriani algorithm [7]:

1. Do the word checking process that you want to do the stem process in the base word database, if the word is found in the database then the algorithm process stops and takes the assumption that the word is the basic word that has been eliminated from various affixes.
2. Doing the process of removing Infection Suffixes ("-lah", "-lah", "-my", "-mu", or "his"). If it is in the form of particles ("-lah", "-lah", "-ah" or "-pun") then this step is repeated again to remove the Positive Pronouns ("my", "your", or "his")) if there are.
3. Dispose of Derivation Suffixes ("-i", "-an" or "-kan"), then check the word in the base word database, if the word is found then the algorithm stops, if not then go to step 3a:
   a. If "-an" has been deleted and the last letter of the word is "-k", then "-k" has also been deleted. If the word is found in the dictionary the algorithm stops.
   b. If not found then do step 3b.Deleted endings ("-i", "-an" or "-kan") are returned, continuing to step 4
4. Eliminating Derivation Suffixes {"di -", "to -", "se -", "me -", "be -", "pe", "te-"} with a maximum of 3 repetitions:
   a. Step 4 stops if:
      1) A forbidden combination of prefix and suffix occurs.
      2) The prefix detected now is the same as the prefix that was previously removed.
      3) Three prefixes have been removed.

   b. Identifying and eliminating prefix types. The prefix has various types:
      1) Standard: "di-", "to", "se" which can be immediately removed from the word.
      2) Complex: "me", "be-", "pe", "te-" are prefix types that can morphology according to the basic words that follow. Therefore, use the rules in Table 1.4 to get the right beheaded.

   c. Search for words that have been omitted in the base word database if they are not found, step 4 is repeated. If found the algorithm is stopped.

5. Doing the process of Recoding or changing the words you want to do the stem process or adding characters in accordance with the method in Indonesian.
6. If all the steps have been done and the input of the word is still not successful then the word is assumed to be the basic word and the process of this algorithm is complete.

F. Stopwords

In the stopwords process, words including prepositions, conjunctions and the like will be removed, this is because in order to transform a feature that is N-Gram and also eliminate data noise so computing is faster.

G. Bag of concept

Feature representation method with the concept of bag of concepts (BOC) is a new development of the concept of preexisting transformation that is using the concept of bag of word model. This form of feature representation takes focus on

a meaning contained in a word and is a combination of several words contained in a document that has the same meaning or meaning.

What needs to be done to implement this feature transformation concept is to do the sum of each existing vector value derived from the calculation of the number of words that appear from each document.

Merging words that have the same meaning or BOC concept according to Täckström (2005) has proven to be implemented well in the information retrieval system even though this concept can be said to be quite simple. However, this concept has a weakness that is the same as the concept of Bag of Word (BOW) model in which the contextual relationships or relationships that exist in a word are not taken into account at all.

H. N-Gram

N-Gram is a new word that results from the cutting technique of a longer string. A characteristic possessed by an N-Gram is the existence of several words that overlap with other words in accordance with the order of the order of the words in the sentence (Permadi, 2008).

According to Hanafi, Whidiana and Dayawati (2009) using the N-Gram method is a very simple approach in carrying out a process of categorizing texts and documents as well as the advantages provided by the N-Gram method is that this method is not too sensitive to the errors in writing contained in a document that will be categorized, but the weakness in using N-Gram to be a feature is a drastic increase in the number of features in the existing document matrix.

The N-Gram implementation is not only based on characters per letter unit but can also be based on word units. The N Index on N-Gram gives a different representation, if the size of N = 1 is called unigram, if the size of N = 2 is called bigram and if the size of N = 3 is called trigram and so on.

III. DISCUSSION AND RESULT

A. Database Design

In the figure below you can see the relational database design.

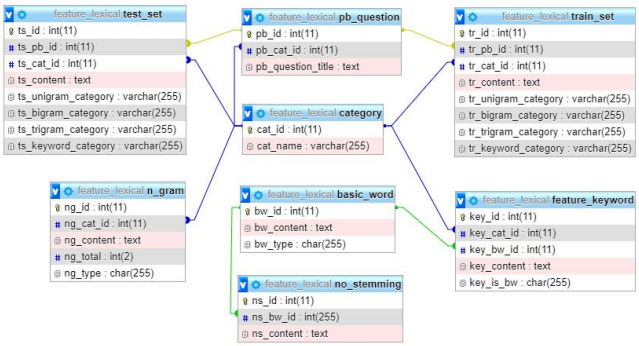

Fig. 3. Relational database

The following detailed tables were created during the implementation of the question category automation application:

a. Table: basic_word
   This table provides basic data on Indonesian words to be used as a reference in the word derivation process.
b. Table: no_stemming
   This table allows you to store all the words that the stemming process has performed with the aim that if there is new data when a word has been stemming, it is not necessary to repeat the existing process and to take only the basic words that are formed in this table.
c. Table: feature_keyword
   This table is used to host several keywords from each of the 10 categories that have been provided in order to implement the word bag of concept in function representations in the document matrix in the function extraction process.

d. Table: n_gram
This table is used to accommodate various word sequences that are included in the n-gram category, both Unigram, Bigram and Trigram. A collection of words comprising n-grams is obtained from the pb_question table, then a process of creating root and keyword is carried out, then the process of forming unigram, bigram and trigram.

e. Table: category
This table is used to provide different categories that are used on the opin.id website and also in the implementation of question categorization.

f. Table: pb_question
This table serves as the main data that provides questions that have a category to use as the data model and questions that have no category to use as the test data to classify.

g. Table: train_set
This table is used to store the data that was performed in the stemming, keyword and entity extraction processes so that train_set is processed in a data model, so if there is any new data to give a category, there is no need to create a data model from the initial stages of preprocessing and data extraction.

h. Table: tes_set
This table is used to answer questions that do not have a temporary category and the extraction, stop words, and entity extraction processes were successful for the purpose that if the question category automation function will be executed in a certain period of time, the question is first recorded so as not to preprocess the data and extract the functionalities again.

B. Stemming process

The stemming process can be seen in the following flow chart.

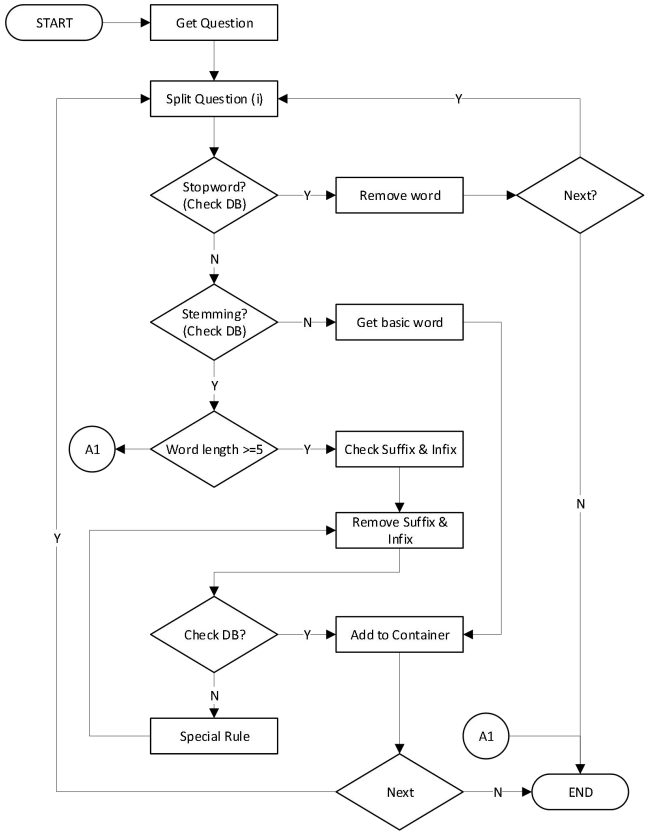

Fig. 4. Stemming flow chat

The following is an explanation of the stemming process:

1. Retrieve the data from the pb_question table.
2. Ask divided questions which are useful for obtaining word units since the derivation process is carried out per word unit.
3. Check the word *i* in the stop words table if the word is included, then the next process is to delete the word and then go back to the beginning to check the word split according to questions, if the word is not in the table, then the following The process is executed.
4. Checking the no_stemming table, if the split word question *i* is in the table, then the next step is to take the basic word and add it to a container, then check if the number of divided words has been exhausted if you do not return to the next division word question step, if the division word in question *i* is not in the table, the following process is carried out.
5. Check the number of letters of the question word divided *i* at least 5.
6. Check the prefix and suffix of the word division question. Then remove the prefix and affix from the word.
7. Check the basic word database if there is a crossword question in the database, then the following process is performed; otherwise, a special rule applies, that is, words that have a fusion.
8. Place the word that was made stemming in the container then check the number of matrices in the divided question if it was used or not, if not, go back to step (3), if it does not exist , then stemming process is complete and produces a return value in the form of words collected in the container.

C. Implementation of feature extraction

Feature extraction is a method used to find information and functionality from an object that is used as a comparison between the data stream used as a data model with the news data that the categorization process you want to do.

The process of extracting functionality in the case of the automation of question categories applies the bag of concept, which is a development of the representation of the bag of word function. The number of features used in the implementation of question category automation is 40 lexical features which are divided into different types. The following are the types of functionality used in implementing question category automation:

1. Groups of keywords in each category
2. Unigram groups of each category
3. Bigram groups in each category.
4. Groups of trigrams in each category.

The bag of concept is a combination of several words which have the same meaning and the implementation used in the transformation of the characteristics has a way of calculating the number of occurrences of the words included in each group of characteristics of each category.

The reason for using lexical functionality in the implementation of question category automation is that the questions that will be used as data models and the questions that will be used as test data will be assigned a category when asked. reading preprocessing data (stemming and stopword) does not become a complete sentence when you use semantic or syntactic characteristics which pay attention to the meaning that emerges from the relationship between existing words and tend to reduce the level of precision considerably so that it is less relevant to be used in industry.

D. Testing

1) Combination of features and classifier test

These are the results of a comparison of combinations of characteristics in the formation of a data model using folds of cross validation = 10 and the example of classifier used is Naïve Bayes, LibSVM, J48 Tree:

TABLE I
RESULTS OF COMBINATION OF CHARACTERISTIC AND CLASSIFIER

Combination Feature	Classifier		
	Naïve Bayes	J48 Tree	LibSVM
UB	89,39	96.04	93.37
UT	80,81	91.30	88.89

UK	41.79	51.92	62.72
UBT	89.23	96.21	94.84
UBTK	79.79	91.02	86.18
UBK	89.48	96.04	92.70
BT	90.12	96.21	95.83
BK	90.43	96.41	95.00
BTK	90.26	**96.50**	95.07
TK	82.36	91.86	90.71
UTK	81.32	91.04	88.60

2) Test a combination of features and classifier with the selection of attributes

The following is the result of the comparison of the combinations of entities in the formation of the data model using folds of validation = 10, the selection of attributes and the examples of classifiers used are Naïve Bayes, LibSVM, J48 Tree :

TABLE I
RESULTS OF COMBINATION OF SELECTION OF CHARACTERISTIC AND CLASSIFIER ATTRIBUTES

Combination Feature	Classifier		
	Naïve Bayes	J48 Tree	LibSVM
UB	89,56	95.78	95.94
UT	82.09	91.09	91.28
UK	44.35	44.48	53.66
UBT	88.80	95.92	95.88
UBTK	82.10	91.10	91.30
UBK	89.56	95.78	95.94
BT	88.80	95.92	95.88
BK	88.28	94.44	94.76
BTK	88.80	95.91	95.88
TK	82.09	91.09	91.28
UTK	82.10	91.10	91.28

Information:
UB = Unigram Bigram
UT = Unigram Trigram
United Kingdom = keyword Unigram
UBT = Unigram Bigram Trigram
UBTK = Unigram Bigram Trigram Keyword
UBK = keyword Unigram Bigram
BT = Bigram Trigram
BK = keyword Bigram
BTK = Bigram Trigram keyword
TK = keyword trigram
UTK = Unigram Trigram Keyword

IV. CONCLUSIONS

The combination of the lexical test characteristics between Unigram, Bigram, Trigram and Keyword of each category in the implementation of data modeling using cross-validation with 10 times shows that the combination of Bigram Trigram

characteristics and Keyword provides a higher percentage of correctly classified instances compared to the other feature combination which is 96.5% with the J48 tree classifier.

The combination of the lexical characteristics test using the attribute selection function to see which characteristics are most important in the process of automating the question category shows that the combination of the lexical characteristics of Unigram Bigram and d 'Unigram Bigram Keyword when using LibSVM Classifier, it provides a higher percentage than the others which is 95, 94%.

REFERENCES

[1] S. Russell and P. Norvig, *Artificial Intelligence A Modern Approach*. New Jersey: Pearson Education, Inc., 2010.

[2] S. Shalev-Shwartz and S. Ben-David, *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

[3] M. Swamynathan, *Mastering Machine Learning with Python in Six Steps - A Practical Implementation Guide to Predictive Data Analytics Using Python*. Bangalore: Apress, 2017.

[4] A. Smola and S. V. . Vishwanathan, *Introduction to Machine Learning*. New York: Cambridge University Press, 2008.

[5] A. Chopra, A. Prashar, and C. Sain, "Natural Language Processing," *Int. J. Technol. Enhanc. Emerg. Eng. Reserach*, vol. 1, no. 4, pp. 131–134, 2013

[6] M. Ikonomakis, S. Kotsiantis, and V. Tampakas, "Text Classification Using Machine Learning Techniques," *WSEAS Trans. Comput.*, vol. 4, no. 8, pp. 966–974, 2005,

[7] M. Adriani, J. Asian, B. Nazief, H. E. Williams, and S. M. M. Tahaghoghi, "Stemming Indonesian: A Confix-Stripping Approach," *Conf. Res. Pract. Inf. Technol. Ser.*, vol. 38, no. September 2018, pp. 307–314, 2007