Effect Verb Extraction on Crime Traditional Cluster

Qusay Bsoul, Juhana Salim and Lailatul Qadri Zakaria

Knowledge Technology Research Group, Center for Artificial Intelligence Technology, Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia 43600 Bangi, Selangor Darul Ehsan, Malaysia

Abstract: Natural language processing (NLP) is a theory-motivated range of computational techniques for the automatic analysis and representation of human language. NLP used in text clustering for extraction the most important terms. The extraction research has evolved from period the Bag of Word to N-gram to part of sentence such as nouns and verbs, where the verbs terms play important role to differentiation each cluster about other. Still this method of extraction need to evaluate the effective on text clustering. This paper draws on use verbs extraction for text clustering using two datasets namely; real datasets about crime and industrial datasets using benchmark datasets. We identified the limitation of traditional clusters such as k-means on the initial centroids, for that's, we generated two new way of choose initial centroids to make the comparison more confident than the previous work.

Key words: Text Clustering • Extractions • Traditional Cluster • K-Means • Real Dataset

INTRODUCTION

Owing to Natural language processing (NLP), extraction method that has mainly been based on algorithms depending on the textual representation of web pages is introduced. Although, such algorithms have very limited capabilities when it comes to interpreting sentences and extracting meaningful information [1], they can efficiently retrieve texts, split them into parts, check the spelling and count the number of words. Text Clustering is considered as one of the most commonly used methods in detecting topics/events or types of documents [2]. It is a method which cluster documents that involves three main processes [3]. The first process is pre-processing of documents is related to any types of document such as Finances, medical and Crimes, which removes unimportant words and symbols from the document. The second process is concerned with extracting from documents to extract the most important information from the text such as Bag of Word. The last process is traditional cluster algorithms. However, this work focuses on evaluate the verb as extraction process and take the Crime case as a real dataset in addition to the industrial datasets as benchmark datasets.

Text clustering algorithms are usually relying on the BOW (Bag of Words) as syntactic extraction using traditional clusters. The BOW cannot accurately represent the most important terms in documents [4], though it take all terms after remove stop-words list and root of the words called stemming. It is also worth mentioning that the textual data have become a diversity of vocabulary owing to the rapid growth of text documents. Besides, they are high-dimensional. Therefore, to correctly represent the theme of documents, improve clustering performance and ideally process data with a small size, text clustering techniques are highly required. Besides, a part of sentence such as nouns have been recently adopted. For instance, to improve the quality of text clustering as one of the most widely used thesauruses for English, has been extensively used [5-14]. The verbs of sentence play important role for distinction between groups of cluster, where the Romanian nation described accurately the definition of ‘crime’ by two concepts ‘actus-reus’ (harmful or awful 'verbs') and ‘mens-rea’ (awful or bad intention 'nouns') [15]. This is method of extraction need to evaluate their performance on text clustering.
This paper is organized as follows: The second section presents the most related work on extraction methods. The third section show the effect verbs as extraction on the real and benchmark datasets. The fourth section generate two new way of choosing initial centroids. The last section will provide the conclusion with future work.

**Review of Extraction Terms:** Recently, electronic systems for reporting crimes have taken the place of the traditional manual reports written on paper in most police departments. Such more advanced or technology-based crime reports include different kinds of information being categorized into types of crime, date/time, location etc. [16]. The role of this categorization is to distinguish between reports of crimes such as ‘theft, drugs, murder’. These categories in crime document are represented by a set of terms which are called features. Extracting the features from the indexed words is considered as a real challenge. This is because more than one feature or combinations of features are considered in this stage. Several researchers have focused on extracting information from the terms or ‘features’ indicating certain crime by employing Name Entity [17], Bag Of Word [18], N-Gram [19], Frequent Word and Frequent Word Meaning [20]. In order to make the extractions step better and more effective, researchers used Concept Weighting Ontology [21], Word Net Lexical Ontology [22] Core Semantic Features [4] Semantic Relationships [23], Word-Net [7], Semantic Method [24], Semantic Preserving [25], Verb-Centric Approach [26] and Pronominal Anaphoric Resolution (RAP) [27].

However, the nature of text data often represents a challenging issue due to the high dimension and ambiguous/overlapping word senses. Previous studies have proposed and developed various term extraction methods to address high dimension of terms such as BOW, N-gram and NER, those methods called syntactic extraction, on one hand and overlapping word senses like semantic words by word-net, latent semantic index and other methods called semantics extraction. In this paper, we initially discussed the syntactic extraction and why researchers proceed to develop and use the semantic extractions.

In syntactic extraction, Zhiwei et al. [3] compared between bag of word and name entity by using Probabilistic Model Clustering. Their findings revealed that the results gained through using the name entity approach were better and more effective than those results generated from the data using bag of word. Yanjunli et al. [20] used Frequent Word and Frequent Word Meaning to compare them with bag of words using Bisecting k-means. Their result showed that Frequent Word and Frequent Word Meaning were better than bag of word.

In evaluating the graphical interface of Document Clustering, Masnizah et al. [28] distinguished between Name Entity and bag of word by using single pass clustering. Their findings revealed that the name entity approach is better than bag of word. Using similar method of syntactic extraction on extract nouns, Fodeh et al. [4] extracted the nouns and compared with all terms extracted on diverse dataset using Spherical k-means to cluster. They found the noun terms were better in all experiments than extracting all terms after pre-processing. In contrast, they did not compare with the baseline of BOW. Al-Shamari & Lin [29] extracted nouns and verbs by using a new method called lemmatization algorithm, which is an idea to extract information from sentences based on certain rules of the sentence. They generated a new technique involving lemmatization algorithm for Arabic Information Retrieval. It aimed to extract nouns and verbs from Arabic documents based on the preposition words as well as some rules related to other linguistic elements such as the definite article “the”. It was reported that this algorithm is better than BOW as it is effective at catching important words from two different lists of prepositions; one including proceeding verbs and one including proceeding nouns.

Sharma et al. [26] proposed an extraction algorithm for extracting verb-centric relationship using Naïve Bayesian classifier. By examining a sentence from biomedical text, their algorithm could identify if it was a relationship bearing sentence or not and then extracted the relationship depicting phrase from the sentence. On other hand, Qasim et al. [30] applied pronominal anaphoric resolution (RAP) by identifying the missing prepositions in the qualified sentences to increase proposition extraction recall using Affinity Propagation cluster and 65 sample documents as data-set. The researchers were able to extract the participating entities which involved around the relationship depicting phrase. Their algorithm is capable of handling missing, incomplete and conjoining entity issues involved in extraction of participating entities. The results showed that this algorithm gained a balanced precision from 0.86 to ~0.95 and a recall form 0.88 to ~0.92 based on their evaluation of three biomedical data sets. On the other
hand, extracted terms as syntactic will hide the core meaning i.e. most of the sense, the hide of meaning will make overlapping between the different terms such as bank, where the term of bank have ten meaning such as company of financial or slope land. This is overlapping in meaning lead to main problem related to syntactic extraction. Though semantic can overcome the weakness of syntactic extraction [31], semantic still has drawbacks, by expand or augment the feature space without necessarily enhancing the clustering performance [4-9] and other drawback related to the number of features may reach hundreds or thousands using the sense of semantic.

To enhance the clustering accuracy with a reduced number of terms, extracting a subset of the disambiguated terms with their relations (known as the core semantic features) that are “cluster-aware” is highly appreciated, Zheng et al. [23] combined detection of noun phrases with the use of WordNet as background knowledge to explore better ways of representing documents semantically for clustering using three divers clusters namely; k-means, bisecting K-means and Hierarchical Agglomerative Clustering. Based on noun phrases as well as single-term analysis, they exploited different document representation methods to analyze the effectiveness of hypernymy, hyponymy, holonymy and meronymy and they used Reuters-21578 as data-set. The results showed that the best method is hypernymy without comparing with the syntactic extraction. Chen et al. [7] used hypernyms of WordNet as proposed to enhance document clustering using Fuzzy-based Multi-label Document Clustering. In their experiment, they used many datasets benchmark and their findings showed that the used hypernyms seemed better than without using the hypernyms. The researchers could increase the accuracy and effectiveness of text mining, but one weakness of their work is concerned with reducing the dimensionality of terms. On the other hand, Fodeh and Tan [4] carried out an experiment using Spherical k-means on semantic features in which all the polysemous and synonymous nouns were extracted from the documents and a unique approach that was capable of permitting them to measure the information gain and disambiguate these nouns in an unsupervised learning setting. The purpose of developing this approach was to identify the core subset of semantic features representing a text corpus. Thus, based on this experiment, the results revealed that by employing core semantic features for clustering, it is more possible to reduce the number of features by 90% or more. At the same time, it is possible to produce clusters that capture the major themes in a text corpus, but the syntactic extraction better than semantic extraction on nouns words.

According to Hmway and Thi [21] and Gharib et al. [22] when they compared Concept Weighting ontology and Word Net Lexical ontology with Bag Of Word using k-means and Self Organizing Map Clusters, the performance of Concept Weighting ontology and Word Net Lexical ontology was much better than bag of word.

As synthesized in the literature review on the extraction process, the semantic still have ambiguity problem on text clustering. Additionally, the empirical results of the previous research have been inconclusive. In particular, some researchers have shown in their experiments that using the concepts appear to give better results [9], but the improvement is not always significant [4-9]. On the other hand, other researchers have reported that the ontological concepts add no value and sometimes impair the performance of document clustering [4]. Where still in the syntactic extraction have a method of extract verb words but did not evaluated before and as we mentioned before in the first section, the verb words have important terms in the text and one of concept support this method is the concept of crime by romanian nation they descript the crime by two main aspects 'nouns and verbs'. However, as shown in the previous researches, they used traditional cluster to evaluate the extraction methods they proposed. The main limitation related to traditional cluster is the initial centroids, where the initial centroids effect on the performance of extraction as mentioned by Bsoul et al. [32], especially, when the extraction methods on text clustering are being compared. The next section evaluated three extraction methods. This paper will evaluate the extraction methods using the real crime dataset and benchmark datasets. Further, two new ways of choosing initial centroids will be used to make fair comparison. It should be noted that one major difficulty in comparing the previously published results is their lack of uniformity, especially in terms of the benchmark data and the baseline algorithms of cluster used.

**Effect of Verb Extraction on Crime Clustering:** The clustering performance may be impacted by some non-descriptive and redundant terms entailed in an ordinary document. Hence, the documents must include a pre-process to get rid of these un-informative terms. The pre-processing steps which are words considered not to convey any meaning, including pronouns, determinants
Table 1: The result of three extraction methods on benchmark and real datasets

<table>
<thead>
<tr>
<th>Datasets</th>
<th>BOW # Term</th>
<th>BOW F-measure</th>
<th>Nouns # Terms</th>
<th>Nouns F-measure</th>
<th>Verbs # Term</th>
<th>Verbs F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>10 types of crime</td>
<td>15600</td>
<td>0.679</td>
<td>6154</td>
<td>0.732</td>
<td>3227</td>
<td>0.723</td>
</tr>
<tr>
<td>6 events of crime</td>
<td>3863</td>
<td>0.593</td>
<td>2376</td>
<td>0.594</td>
<td>1485</td>
<td>0.525</td>
</tr>
<tr>
<td>Routers</td>
<td>12152</td>
<td>0.756</td>
<td>5392</td>
<td>0.736</td>
<td>2931</td>
<td>0.783</td>
</tr>
<tr>
<td>20 Newsgroup</td>
<td>27211</td>
<td>0.671</td>
<td>9935</td>
<td>0.620</td>
<td>4398</td>
<td>0.683</td>
</tr>
</tbody>
</table>

DSB: Data Set Benchmark
DSR: Data Set Real
*best result: underline and bold
*2ed best result: bold

and numbers. In addition, a standard list of 571 stop words is investigated [Stop Word List, http://www.lextek.com/manuals/onix/stopwords2.html] and then deleted from the documents. Besides, various morphological forms can be provided to words with the same meaning. They are normalized into a common root-form, called word stemming, in order to capture their similarity. To conduct the word stemming in this research, the Porter stemmer [33] is adopted. Eventually, in the crime documents, the nouns as well as the verbs are identified and used as terms in this study. It is also worth mentioning that to identify whether a word potentially could be a noun or a verb, Word-Net is adopted in this study by checking if the word is identified in the Word-Net as noun database or as verb database. Therefore, the feature vector for each document will entail the set of stemmed nouns and verbs that resulted from pre-processing. Moreover, to measure the weights of terms and concepts, we need to implement the tfidf (term frequency-inverted document frequency). Finally, documents are represented as feature vectors using the weights of terms concepts, which are then adopted for clustering.

In order to evaluate the effectiveness of the Word-Net-based (noun, verbs) identification method, this research used the samples generated from two real datasets of crime collected from Bernama news. Then, in addition to the benchmark datasets, [34] the researcher collected the Reuters-21578 and 20newsgroups available via a link (http://people.csail.mit.edu/jrennie/20Newsgroups). This research uses the k-means as the clustering algorithm with random initial centroids for each cluster and takes the average of 20 times independent runs. Table 1 shows a comparison of clustering performance between using Bag of word (BOW) as features versus using only the stemmed nouns (as identified by Word-Net) as features and using only the stemmed verbs as identified by Word-Net as features. This work used BOW as the baseline method, which improved 2 out of 4 in text clustering after best result. However, most of previous researches compared the nouns with features without stemming. On the other hand, this research compared the nouns as features with BOW. Additionally, the previous research did not observe the importance of verbs as extraction terms on text clustering, as shown in Table 1. The extraction verbs as terms have been conclusive for the importance of those extracted for clustering where 2 out of 4 are important in clustering and shown the important of used verbs as extraction, once on reduce the number of terms for all datasets used and once on the performance it was satisfied. Still, the performance has ambiguity issue. In particular, some of datasets attained good performance with BOW, while others with Nouns as terms, whereas others with Verbs as terms. All of these are correct because in this experiment, observe that the k-means performance is dependent on the initial centroids. As such, the crime document clustering needs to study the weather of k-means as shown in the following section by specify the initial centroids. In this research proposes a new way of choose initial centroids based on extraction used to show the effect cluster used on extraction as show in next section.

Comparison of Extraction Using Two New Ways: We generated new ways to evaluate the extractions methods: one way is by using the same centroids used by nouns and the other way is by using the centroids generated when using the verbs as extraction. However, it should be noted that these ways of choosing the initial centroids did not simulate the actual of text clustering; rather, it shows the problem related to the cluster algorithms used by previous researchers. Thus, if the difference between ways exists, this will prove this work claim. Figure 1 to 3 shows the comparison between each new way. As shown, the Verbs as extraction better than others ways are used. On the other hand, the Nouns as extraction better than others ways when are used. As such, the performance of
text clustering is still inconclusive. Thus, based on these experiments, two points can be concluded. First point is related to the comparison of the previous researches, where based on the current experiments, the comparison is not fair. Second point is related to the problem on traditional clustering. Clearly, the analysis in this section suggests that such comparison may not be sufficient because one could achieve quite significant improvement in clustering by using their initial centroids as for Nouns or as for Verbs. For that, the different results got it between previous researchers it is correct, where some of these researches show that semantic is better than syntactic, while other researches show that semantic is worse than syntactic.

**CONCLUSION**

The aim of this paper is to evaluate the verb extraction on traditional cluster, where the Romanian nation described accurately the definition of ‘crime’ by two concepts ‘‘actus-reus’ (harmful or awful ‘verbs’) and ‘mens-rea’ (awful or bad intention ‘nouns’) [15]. So that, the verbs word important to differentiation each group about other. In achieving this method of extraction, we evaluate the verb extraction on text Clustering using two datasets namely; report of crime as real dataset and benchmark datasets. In this paper the limitations related to traditional cluster and identification the limitation by generate new two ways of choosing initial centroids the first way by initial centroids of BOW and other way by verbs initial centroids as shown in our experiments. We conducted our experiments to prove the problems identified from previous works and showed the new problem that had effected traditional cluster on extraction used. Based on our new problem, we suggest some of future work to make fair in comparison of extraction methods. Firstly, we suggest to employ meta-heuristic as cluster because meta-heuristic less dependent on initial centroids as traditional cluster ‘k-means’. Secondly, we suggest to combine nouns extraction and verbs extraction to increase the uniqueness of terms, in addition we suggest to evaluate the combine extraction on semantic sense. Besides, we look forward that our suggestion of crime text clustering enhances the performance and the effectiveness.

**REFERENCES**


