가변적 클러스터 개수에 대한 문서군집화 평가방법

조태호
School of Information Technology and Engineering, University of Ottawa
tjo018@naver.com

The Evaluation Measure of Text Clustering for the Variable Number of Clusters

Taeho Jo
School of Information Technology and Engineering, University of Ottawa

요 약

This study proposes an innovative measure for evaluating the performance of text clustering. In using K-means algorithm and Kohonen Networks for text clustering, the number clusters is fixed initially by configuring it as their parameter, while in using single pass algorithm for text clustering, the number of clusters is not predictable. Using labeled documents, the result of text clustering using K-means algorithm or Kohonen Network is able to be evaluated by setting the number of clusters as the number of the given target categories, mapping each cluster to a target category, and using the evaluation measures of text. But in using single pass algorithm, if the number of clusters is different from the number of target categories, such measures are useless for evaluating the result of text clustering. This study proposes an evaluation measure of text clustering based on intra-cluster similarity and inter-cluster similarity, what is called CI (Clustering Index) in this article.

1. Introduction

Text clustering refers to the process of partitioning a collection of documents into several sub-collections of documents based on their similarity in their contents. In the result of text clustering, each sub-collection is called a cluster and includes similar documents in their contents. The desirable principle of text clustering is that documents should be similar as ones within their same cluster and different from ones in their different clusters, in their contents. Text clustering is important tool for organizing documents automatically based on their contents. The organization of documents is necessary to manage documents efficiently for any textual information system. For example, web documents, such as HTML, XML, and SGML, need to be organized for the better web service and emails should be organized based on their contents for the easy access to them. Unsupervised learning algorithms, such as k means algorithm, single pass algorithm, Kohonen Networks, and NTSO, were applied to text clustering as its approaches [1][2][3][4][5]. The evaluation of their performance of text clustering should be performed based on its principle.

Evaluation measures of text categorization, such as accuracy, recall, precision, and F1 measure, were used to evaluate the performance of text clustering in the previous research on text clustering [1][2]. The accuracy is the rate of correctly classified documents to all of documents in the given test bed. This measure is the simplest evaluation measure in classification problems including text categorization, and applicable directly to multi-classification problems. But note that recall, precision, and F1 measure are applicable directly only to binary classification problems. To evaluate the performance of classification using them, the given problem should be decomposed into binary classification problems. In the multi-classification problem, each class corresponds to a binary classification problem, where the positive class indicates “belonging to the class” and the negative class indicates “not belonging to the class”. These evaluation measures focus on only positive class in each binary classification. In text categorization, recall refers to the rate of correctly classified positive documents to all of the true positive documents, precision refers to the rate of correctly classified positive documents to all of the classified positive examples, and F1 measure is the combined value of recall and precision using the equation (1), as follows.

$$F1-measure=\frac{2\times recall\times precision}{recall+precision} \quad (1)$$

The previous research on text clustering proposed and evaluated state of art approaches to text clustering using evaluation measures of text categorization. In 1998, O. Zamir and O. Etzioni proposed suffix tree algorithm as an approach to text clustering and evaluated it using precision. They showed that suffix tree algorithm has higher precision than single pass algorithm and k means algorithm in text clustering [6]. In 1998, S. Kaski and his colleagues proposed a text clustering system, called WEBSOM, where Kohonen Networks are applied as the approach to text clustering [3]. Without evaluating their approach with its comparison with other approaches, they demonstrated the visual result of the system, WEBSOM. In 2000, T. Kohonen and his colleagues revised the system, WEBSOM, to improve its speed to the massive collection of documents by modifying data structures of documents [4]. Although the revised version of WEBSOM is improved even ten times in its speed, both its previous version and its revised version are evaluated using accuracy. In 2000, V. Hatzivassiloglou and his colleagues applied several clustering algorithms, such as single link algorithm, complete link algorithm, group-wise average, and single pass algorithm, to text clustering with and without linguistic features [2]. They evaluated these approaches in these two cases, using linguistic features and not using them, based cost of detection which combines miss and false alarm.

If text categorization based evaluation measures, such as accuracy, F1 measure, and cost of detection are used to evaluate approaches to text clustering in their performance, two conditions are required. For first, all documents in the given test bed should be labelled; they should have their own target categories. In the real world, it is more difficult to obtain labelled document than unlabelled document, and the process of labelling documents follows that of clustering documents in the practical view. The process of preparing labelled documents for the evaluation of approaches to text clustering is time consuming. For second, the number of clusters should be consistent with the number of their target categories. For example, if a series of documents with their same target category is segmented into more than two clusters, text categorization based evaluation measures are useless in that situation.

In 2001, T. Jo proposed an innovative measure of evaluating the result of text clustering [7]. Its advantage over text categorization based evaluation measures is that above two conditions are not required. It does not require the test bed consisting of labelled documents nor the consistency between the number of clusters and the number of their target categories. But it may evaluate the result of text clustering inaccurately, if labelled documents are used as the test bed, because his evaluation measure is computed by analyzing unlabelled documents only lexically. In other words, the similarity between two documents in their same target category may be estimated into its small value. In this case, his proposed evaluation measure is not reliable for evaluating the result of clustering labelled documents.

This study proposes another innovative evaluation measure of text clustering, which is applicable to both labelled and unlabelled documents. In using this evaluation measure of text clustering to labelled documents, the similarity between two documents is given as a binary value, one or zero. If both of them belong to their same target category, their similarity is estimated as one. Otherwise, it is estimated as zero. In using it in unlabelled documents, the similarity between two documents is estimated as a continuous real value between zero and one, using the equations described in the next section by encoding them into one of structured data. Therefore, the proposed evaluation measure solves the problems not only from text categorization based ones but also from the evaluation method proposed in [7].

In the structure of this article, the second section describes the process of evaluating the result of text clustering using the proposed evaluation measure. The third section presents several results of text clustering and their value of their evaluation using the proposed measure in the collection of labelled documents.

2. The Proposed Evaluation Measure

This section describes the evaluation measure of text clustering using labeled documents. The policy of this evaluation is that the better clustering, the higher intra-cluster similarity and the lower inter-cluster similarity. Within the cluster, documents should be as similar as possible, while between clusters, document should be as different possible as possible. This section proposes the evaluation measure reflecting such policy, what is called clustering index, which indicates the rate of intra-cluster similarity to both intra-cluster similarity and inter-cluster similarity. Clustering index is given as a normalized value between zero and one. Its value, 1.0, indicates the completely desirable clustering, where intra-cluster similarity is 1.0 and inter-cluster similarity is 0.0. Its value, 0.0, indicates the completely poor clustering where the average intra-cluster similarity is 0.0, whether the average inter-cluster similarity is any value.

Using a corpus of labeled documents for the evaluation of text clustering, the similarity between two documents is binary value, zero or one. If two documents belong to their same target category, c_t, the similarity between them is 1.0. Otherwise, the similarity is 0.0. The process of computing the similarity between two labeled documents, d_i and d_j is expressed with the equation (2).

$$sim(d_i, d_j) = \begin{cases} 1 & \text{if } d_i, d_j \in c_t \\ 0 & \text{otherwise} \end{cases} \qquad (2)$$

A cluster c_k includes a series of documents and is denoted as a set of documents by $c_k = \{d_{k1}, d_{k2}, \ldots, d_{k|c_k|}\}$. The intra-cluster similarity of the cluster, c_k, σ_k is computed using the equation

(3) and indicates the average similarity of all pairs of different documents included in the cluster, c_k.

$$\sigma_k = \frac{2}{|c_k|(|c_k|-1)}\sum_{i>j} sim(d_{ki}, d_{kj}) \qquad (3)$$

If a series of clusters as the result of text clustering is denoted by $C = \{c_1, c_2, ..., c_{|C|}\}$, the average intra-cluster similarity, $\overline{\sigma}$ is computed using the equation (4), by averaging the intra-cluster similarities of the given clusters.

$$\overline{\sigma} = \frac{1}{|C|}\sum_{k=1}^{|C|}\sigma_k \qquad (4)$$

The inter-cluster similarity between two clusters, c_k and c_l, δ_{kl}, is computed using the equation (5) and indicates the average similarity of all possible pairs of two documents belonging to their different clusters.

$$\delta_{kl} = \frac{1}{|c_k||c_l|}\sum_{i=1}^{|c_k|}\sum_{j=1}^{|c_l|} sim(d_{ki}, d_{lj}) \qquad (5)$$

The average inter-cluster similarity $\overline{\delta}$ is computed using the equation (6), by averaging all possible pairs of different clusters.

$$\overline{\delta} = \frac{2}{|C|(|C|-1)}\sum_{k>l}\delta_{kl} \qquad (6)$$

From the equation (3) to the equation (7), the average intra-cluster similarity, $\overline{\sigma}$ and the average inter-cluster similarity, $\overline{\delta}$, over the given clusters are obtained. Therefore, the clustering index, CI is computed using the equation (7).

$$CI = \frac{\overline{\sigma}^2}{\overline{\sigma} + \overline{\delta}} \qquad (7)$$

The equation (7) shows that a normalized value between zero and one is given in the clustering index. If CI is 1.0, indicates that the average intra-cluster similarity is 1.0 and the overage inter-cluster similarity is 0.0. If the average intra-cluster similarity is 0.0, CI is absolutely 0.0. The equation (7) implies that both intra-cluster similarity and inter-cluster similarity should be considered for evaluating the result of text clustering.

3. Results of Evaluating Text Clustering

There are two experiments using the collection of labeled documents in this section: the consistency and the inconsistency between clusters and their target categories in their number. In the first experiment, the proposed measure is compared with text categorization based evaluation measures: accuracy, recall, precision, and F1 measure. These evaluation measures are compared each other in two cases: the desired clustering where documents are arranged according their target categories and several cases of random clustering where documents are arranged at random with the regardless of their target categories, but the number of their clusters is same to that of their categories.

The collection of labeled documents, which is used in this experiment, includes four hundreds news articles labeled with one of four categories in ASCII text files. The predefined categories in such collection are, "corporate news", "criminal law enforcement", "economical index", and "Internet". This collection was obtained by copying news articles from the web site, www.newspage.com, and pasting them as ASCII text files, individually. Each category includes one hundred news articles, equally.

In this experiment, the number of clusters is set as that of their target categories; four clusters are given. In the desired clustering, each cluster is corresponds to one of their target categories and each document is arranged to its corresponding cluster. In a random clustering, each document is arranged to one of these four clusters at random. By doing this, four sets of random clustering are built. The evaluation measure to each set of text clustering is computed, using the equation (7).

In the desired clustering, the value of the proposed evaluation measure expressed with the equation (7) is 1.0, since the average intra-cluster similarity is 1.0 and the average inter-cluster similarity is 0.0 based on the equation (2). If it is evaluated using text categorization based evaluation measures, accuracy, precision, recall, and F1 measures have 1.0 as their values. Therefore, both the proposed evaluation measure and the text categorization based ones evaluate the result of the desired clustering, identically.

A result of text clustering is presented in the table 2. The number of clusters is identical to that of the target categories of documents, and each cluster is identical to each target category in their number of documents. To apply text categorization based method, each cluster must correspond to one of target categories exclusively. According the majority of each cluster and one to one correspondence, cluster 1, cluster 2, cluster 3, and cluster 4 correspond to corporate news, criminal law enforcement, Internet, and economic index, respectively. Cluster 1, cluster 2, and cluster3 were matched with the target categories according their majority, but cluster 4 was matched with economic index, exceptionally, since each cluster was not allowed to correspond to a redundant category in one to one correspondence. In this condition, all of text categorization based evaluation measures, such as accuracy, recall, precision, and F1 measure, resulted in 0.475 uniformly. In the proposed evaluation method, average intra-cluster similarity is 0.38, using equation (2), (3), and (4) and inter-cluster similarity is 0.1808, using equation (2), (5), and (6). Therefore, the clustering index is 0.2574, using the equation (7).

Table 1. A Result of Clustering News Articles

	cluster 1	cluster 2	cluster 3	cluster 4	Total
corporate news	70	10	10	10	100
criminal law enforcement	15	50	5	30	100
economic index	5	30	40	25	100
Internet	10	10	45	35	100
Total	100	100	100	100	400

The table 2 presents another result of clustering news articles. In this result, cluster 1 and cluster 4 have 150 documents and 50 documents differently from target categories. This leads to difference between recall and precision. According the majority of each cluster and one to one correspondence, cluster 1, cluster 2, cluster 3, and cluster 4 correspond to corporate news, economic index, Internet, and criminal law enforcement, in order to use text categorization based evaluation measures. Accuracy and recall of this result are 0.45 identically. Precision and F1 measure are 0.3665 and 0.4039, respectively. In the proposed evaluation measure, the average intra-cluster similarity is 0.4153 and the average inter-cluster similarity is 0.2054. The clustering index is estimated as 0.2776 indicating that these news articles are clustered better than random clustering, at least. Note that the proposed evaluation measure does not require such correspondence between each cluster and each target category.

Table 2. A Result of Clustering News Articles

	cluster 1	cluster 2	cluster 3	cluster 4	Total
corporate news	70	5	15	10	100
criminal law enforcement	60	40	0	0	100
economic index	15	50	25	10	100
Internet	5	5	60	30	100
Total	150	100	100	50	400

If the number of clusters is not same to that of target categories, text categorization based evaluation measure becomes useless, since the correspondence between clusters and target categories can not be one to one. If the collection of news articles is partitioned into five clusters, where three clusters are exactly same to three of target categories and a particular target category are partitioned into two clusters, the average intra-cluster similarity 1.0, but the average inter-cluster similarity is 0.1. There are ten pairs of clusters among five clusters and one of ten pairs is 1.0; the average inter-cluster similarity is 0.1. Therefore, the clustering index is computed as 0.9090 using the equation (7). On contrary, two target categories may be merged into a cluster. For example, two target categories are same to two clusters in their distribution of documents, but the rest categories are merged into a cluster in this collection of news articles. The average intra-cluster similarity is 0.8324 and the average inter-cluster similarity is 0.0 in this case. Therefore, the clustering index is computed as 0.8324, using the equation (7).

The table 3 presents one of realistic results of text clustering, where the number of clusters is different from that of the target categories of documents, in the second experiment. As mentioned above, text categorization based evaluation measures are not applicable, since these clusters are not able to correspond to these target categories one to one. In this result illustrated in the table 3, the average intra-cluster similarity is 0.3203 and the average inter-cluster similarity is 0.2170. Using the equation (7), the clustering index is 0.1909.

Table 3. A Result of Clustering News Articles

	cluster 1	cluster 2	cluster 3	Total
corporate news	70	20	10	100
criminal law enforcement	30	30	40	100
economic index	40	50	10	100
Internet	10	70	20	100
Total	150	170	80	400

The main title (on the first page) should begin 1-3/8 inches (3.49 cm) from the top edge of the page, centered, and in Times 14-point, boldface type. Capitalize the first letter of nouns, pronouns, verbs, adjectives, and adverbs; do not capitalize articles, coordinate conjunctions, or prepositions (unless the title begins with such a word). Leave two 12-point blank lines after the title.

These two experiments in using labelled documents as test bed for text clustering show that the proposed evaluation method is more suitable for text clustering than the text categorization based evaluation methods with two points. The first point is that text categorization based evaluation methods require the one to one correspondence between clusters and target categories, but the proposed method does not require it. When the number of clusters is same to that of the target categories, each cluster should be matched with a category exclusively. When the number of clusters is different from that of target categories, text categorization based evaluation method is useless. The second point is that text categorization based evaluation measures do not consider the similarities between clusters. This ignores the second principle of text clustering, "documents in different clusters should be different as much as possible". The proposed evaluation measure considers the similarities of documents not only within a particular cluster but also between two different clusters.

4. Conclusion

This paper proposed an innovative evaluation measure of text clustering. This measure underlies the principle, as follows.
The documents within a particular cluster should be as similar as possible, and those between two documents should be as different as possible.
Based on this principle, this study proposed the process of computing the intra-cluster similarity, using the equation (2), (3),

and (4) and the inter-cluster similarity, using the equation (2), (5), and (6). The final evaluation measure of text clustering is computed using these two measures with the equation (7).

When the number of clusters is same to that of target categories, the proposed measure was compared with text categorization based evaluation measures in the previous section. The experiment in that section showed two advantages of the proposed method over text categorization based ones. The first advantage is that each category does not need to be matched with a cluster, in using the proposed evaluation measure of text clustering. Its advantage leads to that it is applicable although the number of clusters is different from that of target categories. The second advantage over text categorization based method is that the proposed evaluation measure considers both intra-cluster similarity and inter-cluster similarity. Text categorization based measures, such as accuracy, recall, precision, and F1 measure, evaluate the result of text clustering based only on intra-cluster similarity.

There is one more advantage of the proposed evaluation measure over text categorization based ones. The advantage is that the proposed measure is applicable even to unlabeled documents, if the process of computing a semantic similarity between two documents is defined. In the real world, it is far easier to obtain unlabeled documents than labeled documents. The assumption underlying in text clustering is that every document is not labeled initially. Therefore, the effort to obtain labeled documents for the evaluation of text clustering is not necessary, in using the proposed evaluation measure.
In the real world, almost every document is labeled with more than one category. In the collection of news articles called Reuter 21578, which is used as a standard test bed for the evaluation of text categorization, each news articles has more than one category. Although overlapping clustering, where a document is allowed to be arranged into more than two clusters, is more practical than exclusive clustering in the real world, the previous research on text clustering focused on exclusive clustering for their easy evaluation. The proposed evaluation measure may be applicable depending on how to define the similarity between documents as expressed in the equation (2).

In the further research, the proposed evaluation method of text clustering is modified to be applicable to the collection of unlabeled documents, the collection documents labeled with more than one category like Reuter 21578, and the hybrid collection of labeled and unlabeled documents. There are several strategies of using the proposed evaluation method to hybrid collections. In a strategy, unlabeled documents are classified with the reference to labeled documents; all of documents are labeled documents. In another strategy, the similarity is computed using the equation (2) if two documents are labeled, and the similarity is computed differently, otherwise. By modifying the proposed evaluation measure to be applicable to the collection of various documents, the flexibility of the proposed evaluation measure is expected to be improved.

References

[1] M. Steinbach and G. Karypis and V. Kumar, A comparison of document clustering techniques, *in the Workshop on Text Mining in SIGKDD*, 2000.

[2] V. Hatzivassiloglou, L. Gravano, and A. Maganti, An Investigation of Linguistic Features and Clustering Algorithms for Topical Document Clustering, *The Proceedings of 23rd SIGIR*, 2000, 224-231.

[3] S. Kaski, T. Honkela, K. Lagus, and T. Kohonen, WEBSOM-Self Organizing Maps of Document Collections, *Neurocomputing, 21*, 1998, 101-117.

[4] T. Kohonen, S. Kaski, K. Lagus, J. Salojarvi, V. Paatero, and A. Saarela, Self Organization of a Massive Document Collection, *IEEE Transaction on Neural Networks, 11* (3), 2000, 574-585.

[5] T. Jo and N. Japkowicz, Text Clustering using NTSO, *The Proceedings of IEEE IJCNN*, 2005, 558-563.

[6] O. Zamir and O. Etzioni, Web Document Clustering: A Feasibility Demonstration, *The Proceedings of SIGIR 98*, 1998, 46-54.

[7] T. Jo, Evlauation Function of Document Clustering based on Term Entropy, *The Proceedings of 2nd International Symposium on Advanced Intelligent System*, 2001, 302-306.

[8] F. Sebastiani, Machine Learning in Automated Text Categorization, *ACM Computing Survey, 34* (1), 2002, 1-47.

[9] T. Jo, NeuroTextCategorizer: A New Model of Neural Network for Text Categorization, *The Proceedings of International Conference of Neural Information Processing 2000*, 2000, 280-285.

[10] T. Jo, Machine Learning based Approach to Text Categorization with Resampling Methods, *The Proceedings of the 8th World Multi-Conference on Systemics, Cybernetics and Informatics*, 2004, 93-98.