CONCEPT AND TERM BASED SIMILARITY MEASURE FOR TEXT CLASSIFICATION AND CLUSTERING

B Sindhiya1* and N Tajunisha1

*Corresponding Author: B Sindhiya email4sini.1@gmail.com

The exploitation of syntactic structures and semantic background knowledge has always been an appealing subject in the context of data mining, text retrieval and information management. The usefulness of this kind of information has been shown most prominently in highly specialized tasks, such as text categorization scenarios. So far, however, additional syntactic or semantic information has been used only individually. In this paper, a new principle approach, the concept and term based similarity measure, which incorporates linguistic and semantic structures, using syntactic dependencies, and semantic background knowledge is proposed. This novel method represents the meaning of texts in a high-dimensional space of concepts derived from WordNet. A number of case studies have been included in the research to demonstrate the various aspects of this framework.

Keywords: Document classification, Document clustering, Similarity measure, Accuracy, Classifiers, Clustering algorithms

INTRODUCTION

Clustering maps the data items into clusters, where clusters are natural grouping of data items based on similarity or probability density methods. Unlike classification and prediction which analyzes class-label data objects, clustering analyzes data objects without class-labels and tries to generate such labels. A similarity measure or similarity function is a real-valued function that quantifies the similarity between two objects.

The similarity measure reflects the degree of closeness or separation of the target objects and should correspond to the characteristics that are believed to distinguish the clusters embedded in the data. Before Clustering, a similarity/distance measure must be determined (Chim and Deng, 2008). Choosing an appropriate similarity measure is also crucial for cluster analysis, especially for a particular type of clustering algorithms.

Text Categorization (TC) is the classification of documents with respect to a set of one or more preexisting categories (Sebastiani, 2002). The classification phase consists of generating a
weighted vector for all categories, then using a
similarity measure to find the closest category.
The similarity measure is used to determine the
degree of resemblance between two vectors. To
achieve reasonable classification results, a
similarity measure should generally respond with
larger values to documents that belong to the
same class and with smaller values otherwise.
During the last decades, a large number of
methods proposed for text categorization were
typically based on the classical Bag-of-Words
model where each term or term stem is an
independent feature.

The existing similarity measure was more
frequently used to assess the similarity between
words. Although the information theoretic similarity
measure results are statistically significant it does
not reduce the dimension of the vector model
(Clinchant S and Gaussier, 2010). Metric
distances such as Euclidean distance are not
appropriate for high dimension and sparse
domains. Due to the ignorance of any relation
between words, the learning algorithms are
restricted to detect patterns in the used
terminology only, while conceptual patterns
remain ignored.

Existing approaches requires performing an
optimization over an entire collection of
documents. Most of these techniques are
computationally expensive.

RELATED WORKS
Similarity measures have been extensively used
in text classification and clustering algorithms.
The spherical k-means algorithm (2007) (http://
web.ist.utl.pt/ acardoso/datasets) adopted the
cosine similarity measure for document
clustering. In this method the unlabeled document
collections are becoming increasingly common
and available. Using words as features, text
documents are often represented as high-
dimensional and sparse vectors. The algorithm
outputs k disjoint clusters each with a concept
vector that is the centroid of the cluster
normalized to have unit Euclidean norm.

Derrick Higgins (2007) adopted a cosine-
based pairwise adaptive similarity measure for
document clustering. Pairwise-adaptive similarity
measure for large high dimensional document
datasets improves the unsupervised clustering
quality and speed compared to the original cosine
reported results of clustering experiments with
clustering algorithms and 12 different text data
sets, and concluded that the objective function
based on cosine similarity “leads to the best
solutions irrespective of the number of clusters
for most of the data sets.”

Daphe Koller and Mehran Sahami (1997)
proposed a divisive information-theoretic feature
clustering algorithm for text classification using
the Kullback-Leibler divergence. High
dimensionality of text can be a deterrent in
applying complex learners such as Support Vector
Machines to the task of text classification

Kullback and Leibler (1951) combined
squared Euclidean distance with relative entropy
in a k-means like clustering algorithm. K means
algorithm introduced recently is specifically
designed to handle unit length document vectors.
(Zoulficar younes, 2003) conclude that the
objective function based on cosine similarity leads
to the best solutions irrespective of the number
of clusters for most of the data sets.

Chim and Deng (2008) performed document
clustering based on the proposed phrase based
similarity measure. The phrase-based document
similarity to compute the pair-wise similarities of documents based on the Suffix Tree Document (STD) model. By mapping each node in the suffix tree of STD model into a unique feature term in the Vector Space Document (VSD) model, the phrase-based document similarity naturally inherits the term tf-idf weighting scheme in computing the document similarity with phrases.

**PROPOSED METHODOLOGY**

In this chapter the proposed concept and term based similarity between the documents is illustrated. The proposed system, also measure the semantic similarity between the terms and concepts with use of wordnet tool and tree tagger tool.

**Concept Based Similarity Measure For Text Processing (CSMTP) Algorithm**

The CSMTP algorithm selects the terms from the testing documents, generates the terms from the document, selects the appropriate feature and calculates the similarity measure based on the term and its respective concepts.

**Csmtp Algorithm**

1. Let $D_1$ and $D_2$ be the testing documents.
2. Let $T_1$ and $T_2$ be the terms from the document $D_1$ and $D_2$.
3. Remove the stopwords $S_{T_1}$ and $S_{T_2}$ from the documents $D_1$ and $D_2$.
4. Let $C_1$ and $C_2$ be two concepts from $T_1$ and $T_2$ respectively where ($T_1$ denotes the first thesaurus and $T_2$ the second).
5. Compute the similarity measure between two concepts, with,

$$SIM(c_i, c_k) = 1 - \frac{\log \left( \frac{\max(|A|,|B|)}{\log(|W|)} \right) - \log \left( |A \cap B| \right)}{\log \left( \min(|A|,|B|) \right)}$$

where $A$ and $B$ are the sets of all articles that link to concepts $c_i$ and $c_k$ respectively and $W$ is the set of all articles. $\max(A,B)$ represents the maximum similarity measure between $A$ and $B$. $\min(A,B)$ represents the minimum similarity measure between $A$ and $B$. The $\log(A \cap B)$ represents the common concepts in $A$ and $B$.

6. Compute the semantic relatedness between term and its candidate concepts in a given document according to the context information

$$Rel(t, c_j | d_j) = \frac{1}{|T|-1} \sum_{t \in \bar{d}_j} \frac{1}{|CS_j|} \sum_{c \in CS_j} SIM(c_i, c_k)$$

where $T$ is the term set of the $j$th document $d_j$, $t$ is a term in $d_j$ except for $t$, $CS_1$ is the candidate concept set related to term $t$.

**Syntactic Representation**

Tf-idf weighting scheme are used in syntactic level to record the syntactic information. Tf–idf, term frequency–inverse document frequency, is a numerical statistic which reflects how important a word is to a document in a collection or corpus. It is often used as a weighting factor in information retrieval and text mining. The tf-idf value increases proportionally to the number of times a word appears in the document, but is offset by the frequency of the word in the corpus, which helps to control for the fact that some words are generally more common than others. Tf–idf is the product of two statistics, term frequency and inverse document frequency. Various ways for determining the exact values of both statistics exist. In the case of the term frequency $tf(t,d)$, the simplest choice is to use the raw frequency of a term in a document, i.e. the number of times that...
Semantic Similarity

Semantic level consists of concepts related to the terms in the syntactic level. These two levels are connected via the semantic correlation between terms and their relevant concepts.

WordNet is used to calculate the ascertain connections among four types of Parts of Speech (POS) - noun, verb, adjective, and adverb. The minimum unit in a WordNet is synset, which represent an exact meaning of a word. It includes the word, its clarification, and its synonyms.

EXPERIMENTAL RESULTS

In this section, the effectiveness of the proposed similarity measure CSMTP is investigated. The investigation is done by applying the CSMTP measure in several text applications, including k-NN based single-label classification (SL-kNN), k-NN based multi-label classification (ML-kNN), k-means clustering (k-means), and hierarchical agglomerative clustering (HAC). The data sets, namely WebKB, Reuters-8 respectively, are used in the experiments presented below.

WebKB

The documents in the WebKB data set are webpages collected by the World Wide Knowledge Base (Web→Kb) project of the CMU text learning group. The documents were manually classified into several different classes. The documents of this data set were not predesignated as training or testing patterns. the datasets can be randomly divided into training and testing subsets.

Reuters-8

Reuters-21578 ModeApt’e Split Text Categorization Test Collection contains thousands of documents collected from Reuters newswire in 1987. The most widely used version is Reuters-21578 ModeApt’e, which contains 90 categories and 12902 documents.

Classification Performance

For WebKB dataset, the randomly selected training documents are used for training/validation and the testing documents are used for testing. For Reuters-8 dataset the predesignated training data are used for training/validation and the predesignated testing data are used for testing. Note that the data for training/validation are separate from the data for testing in each case.

Single-Label Document Classification

In this experiment, we compare the performance of our measureand the others in single-label document classification. The performance is evaluated by the classification accuracy, AC, which compares the predicted label of each document with that provided by the document corpus:

\[
\text{ACCURACY} = \frac{\sum_{i=1}^{n} E(c_i, c_{i1})}{n}
\]

where n is the number of testing documents, and \(c_i\) and \(c_{i1}\) are the target label and the predicted label, respectively, of the ith document. \(E(c_i, c_{i1}) = 1\) if \(c_i = c_{i1}\), and \(E(c_i, c_{i1}) = 0\) otherwise.

Figure 1 shows the classification accuracy obtained by SL-kNN with SMTP and CSMTP similarity measures using different class(k) different k(class) settings, i.e., \(k = 4, 8, 12, 16, 20\), on the training/validation data of WebKB. The figure clearly shows that the single label document classification accuracy obtained using
the proposed CSMTP measure performs high comparing to the SMTP measure.

Figure 1: Classification Accuracy by Sl-Knn On Training/Validation With SMTP & CSMTP Using Different K Settings

Table 1 shows the classification AC(accuracy) obtained by single label classification on the testing data of webkb.

Table 1: Classification Accuracies by Sl–Knn with Different Measures on Testing Data of Webkb

<table>
<thead>
<tr>
<th>K=1</th>
<th>K=3</th>
<th>K=5</th>
<th>K=7</th>
<th>K=9</th>
</tr>
</thead>
<tbody>
<tr>
<td>SMTP</td>
<td>0.9013</td>
<td>0.9191</td>
<td>0.9242</td>
<td>0.9223</td>
</tr>
<tr>
<td>CSMTP</td>
<td>0.9338</td>
<td>0.9411</td>
<td>0.9420</td>
<td>0.9447</td>
</tr>
</tbody>
</table>

Multi Label Classification

Table 2 shows the classification AC(accuracy) obtained by multi label classification on the testing data of webkb.

Table 2: Classification Accuracies by Ml–Knn With Different Measures on Testing Data of Webkb

<table>
<thead>
<tr>
<th>K=1</th>
<th>K=3</th>
<th>K=5</th>
<th>K=7</th>
<th>K=9</th>
</tr>
</thead>
<tbody>
<tr>
<td>SMTP</td>
<td>0.6910</td>
<td>0.6932</td>
<td>0.6965</td>
<td>0.6990</td>
</tr>
<tr>
<td>CSMTP</td>
<td>0.7130</td>
<td>0.7111</td>
<td>0.7114</td>
<td>0.7092</td>
</tr>
</tbody>
</table>

Clustering Performance

For a document corpus with p classes and n documents, remove the class labels and randomly select one-third of the documents for training/validation and the remaining for testing. Note that the data for training/validation are separate from the data for testing.

Kmeans Clustering

In this experiment, the performance of the CSMTP measure in clustering is compared with the SMTP measure. The performance is evaluated by the clustering accuracy, AC, which compares the predicted label of each document with that provided by the document corpus:

$$\text{ACCURACY} = \frac{\sum_{i=1}^{n} \text{most}_{i}}{n}$$
Figure 3 shows the clustering accuracy obtained by kmeans with SMTP and CSMTP similarity measures using different k(cluster) settings, i.e., k=4,8,12,16,20, on the training/validation data of WebKB. The figure clearly shows that the kmeans clustering accuracy obtained using the proposed CSMTP measure performs high comparing to the SMTP measure.

Table 3 shows the clustering AC(accuracy) obtained by kmeans clustering on the testing data of webKB.

<table>
<thead>
<tr>
<th>K=8</th>
<th>K=16</th>
<th>K=24</th>
<th>K=32</th>
</tr>
</thead>
<tbody>
<tr>
<td>SMTP</td>
<td>0.7906</td>
<td>0.8584</td>
<td>0.8692</td>
</tr>
<tr>
<td>CSMTP</td>
<td>0.8450</td>
<td>0.8702</td>
<td>0.8796</td>
</tr>
</tbody>
</table>

Hierarchical Agglomerative Document Clustering Performance

Figure 4 shows the Clustering accuracy obtained by hierarchical clustering with SMTP and CSMTP similarity measures using different k(cluster) settings, i.e., k=4,8,12,16,20, on the training/validation data of WebKB. The figure clearly shows that the hierarchical agglomerative clustering accuracy obtained using the proposed CSMTP measure performs high comparing to the SMTP measure.

Table 4 shows the classification AC(accuracy) obtained by multi label classification on the testing data of webkb.

<table>
<thead>
<tr>
<th>K=1</th>
<th>K=3</th>
<th>K=5</th>
<th>K=7</th>
<th>K=9</th>
</tr>
</thead>
<tbody>
<tr>
<td>SMTP</td>
<td>0.5960</td>
<td>0.5483</td>
<td>0.5367</td>
<td>0.5091</td>
</tr>
<tr>
<td>CSMTP</td>
<td>0.6770</td>
<td>0.6343</td>
<td>0.5552</td>
<td>0.5244</td>
</tr>
</tbody>
</table>

CONCLUSION

The concept and term based model represents document as a two-way model with the aid of WordNet. In the two-way representation model, the term information is represented first, and the concept information is represented second and these levels are connected by the semantic relatedness between terms and concepts. Experimental results on real data sets have shown that the proposed model and classification framework significantly improved the classification and clustering performance by comparing with the existing SMTP(similarity measure for text processing) model.
experiments shows CSMTP (concept and term based similarity measure for text processing) takes less time when running in parallel, less space when running in series and categorization accuracy is high.

**FUTURE WORK**

In future, the work can be focused on the concept mapping and weighting technology to find the better concept vector space for documents, because the better concept-based representation can help to further improve the performance of text classification and clustering framework. A new semantic-based vector space model utilizing the category information can also be exploited. Afterwards, two-way representation model can be extended to three-way model containing term, concept and category information respectively. CTSMTF will also be improved to fit the three-way model and achieve more predominant text classification and clustering performance.

**REFERENCES**


This article can be downloaded from http://www.ijerst.com/currentissue.php