Impact of Term Weighting Schemes on Document Clustering – A Review

G. Hannah Grace and Kalyani Desikan
Department of Mathematics, VIT University, Chennai
hannahgrace.g@vit.ac.in, kalyanidesikan@vit.ac.in

Abstract

Term weighting schemes are used to identify the importance of terms in a document collection and assign weights to them accordingly. Document clustering uses these term weights to identify if documents are similar. In this article, we apply different term weighting schemes to a document corpus and study their impact on document clustering. At first, the given document corpus (DC), is pre-processed using tokenization, stopwords removal and stemming process and then converted to its term document matrix. We have worked with six term weighting schemes viz.: TF, TFIDF, MI, ATC, Okapi, TFICF and obtained clustering solutions using k-means clustering algorithm. In this review paper, we have compared the clustering solutions obtained based on the well-known cluster quality measures: entropy and purity.

Key Words and Phrases: Document Clustering; Term Weighting Schemes; TF; TFIDF

1 Introduction

Document clustering is an automatic grouping of text documents into clusters such that documents within a cluster have high similarity in comparison to one another, but are dissimilar to documents in other clusters. Document clustering is a method to find useful features from the document corpus. It is used to understand the
structure and content of unknown text sets as well as to give new perspectives on familiar ones [1].

Our paper is organized as follows. Section 2 describes the representation of the term document matrix after preprocessing. Section 3 discusses the various term weighting schemes like TF, TFIDF, MI, ATC, Okapi, TFICF and their formulae. Section 4 outlines the experimental results and the analysis based on cluster quality measures. Section 5 concludes the result. In this review paper we have analyzed the various weighting schemes, found the clustering solution and validated the results with cluster quality measures such as entropy and purity.

2 Document preprocessing and Representation of Term Document Matrix

Given a document corpus DC it is pre-processed using techniques like removing blank spaces, numbers, web links, punctuation marks, stop words and stemming [2]. The vector space model [3] is one of the most well known models that is used to represent documents. The main task is to find an appropriate encoding of the feature vector. This provides an efficient quantitative representation of each document [4].

After reducing the sparseness, we formulate the term document matrix. Let \( T = t_1, t_2, ..., t_p \) where \( t_i \) are terms obtained after preprocessing. We then formulate the term document matrix \( X \) as in Figure 1. i.e., \( X = [x_{ij}]_{N \times p} \) is the term document matrix and each document \( d_i \) is a tuple in \( p \)-dimensional space \( R^p \). Each \( x_{ij} \) normally, represents the frequency of term \( t_j \) in document \( d_i \). Apart from term frequency other weighting schemes can also be used. The objective of this paper is to analyse the impact of six term weighting schemes on document clustering.

3 Term Weighting Schemes

Term weighting schemes are important for applications based of texts and hence have been well studied. Term weighting [5] is a numerical representation of a collection of documents described by
'p' terms and 'N' documents from a document corpus DC. This $X = [x_{ij}]_{N \times p}$ is a sparse matrix called the term-document matrix whose rows correspond to documents and columns correspond to the stemmed terms appearing in the documents. There are three different factors that determine term weighting schemes [6] : local, global and normalization.

### 3.1 TF weighting Scheme

Term frequency (TF) is the simplest measure to weight each term in a document. Term frequency (TF) measures how frequently a term occurs in a document. In this weighting scheme, a term document matrix comprises of rows corresponding to documents in the collection and columns corresponding to terms. The elements of this matrix show which document contains which term and how many times they appear in each of the documents.

### 3.2 TFIDF weighting scheme

TFIDF is term frequency - inverse document frequency weighting scheme which is commonly used. This reflects how important a word is to a document in a collection or corpus. The TFIDF value increases proportional to the number of times a word appears in the document using the frequency of the word in the corpus. The
The basic form of TFIDF is given by

$$w_{ij} = x_{ij} \log \frac{N}{n_j}$$

where $w_{ij}$ is the weight of term $j$ in document $i$; $x_{ij}$ is the number of occurrences of term $j$ in document $i$ (TF); $N$ is the total number of documents in the document collection; and $n_j$ is the number of documents in which term $j$ occurs at least once. $\frac{n_j}{N}$ is often referred to as the document frequency (DF) of term $j$ and naturally, $\frac{N}{n_j}$ is called the inverse document frequency (IDF) of term $j$. IDF is the most frequently used collection frequency factor. The reason for incorporating IDF is because of its simplicity and robustness. But it has its own drawbacks. IDF weight is based on the intuition that words that do not occur frequently in a collection tend to be more informative than the words that appear frequently across many documents.

### 3.3 MI weighting scheme

The mutual information method (MI) is a supervised method that works by computing the mutual information between a given term and a document class. This provides a measure of how much information the term can tell about a given class of documents. Low mutual information suggests that the term has a low discrimination power and hence it has to be removed [8]. The weighting scheme is given by,

$$w_{ij} = \frac{x_{ij}}{N} \frac{N}{\sum_{i=1}^{N} x_{ij} \sum_{j=1}^{p} x_{ij}}$$

where $N$ is the number of documents in the document collection, $p$ is the number of terms and $x_{ij}$ is the term document frequency.

### 3.4 Okapi weighting scheme

Okapi[7] is one of the best-known term weighting schemes. It originated from the probabilistic relevance model, rather than the vector space model. But Okapi has a lot in common with the TFIDF
weighting scheme which is based on the vector space model. Both use term frequency, inverse document frequency and normalisation.

Okapi term weighting scheme is given as follows:

\[
  w_{ij} = \left[ \frac{x_{ij}}{0.5 + 1.5 \left( \frac{d_l}{avg.d_l} + x_{ij} \right)} \right] \log \left( \frac{N - n_j + 0.5}{x_{ij} + 0.5} \right)
\]

where \(x_{ij}\) is the term document frequency, \(N\) is the total number of documents in the collection, \(n_j\) is the document frequency, i.e., the number of documents in which term \(j\) occurs, \(d_l\) is the document length i.e., the number of terms in a particular document and \(avg.d_l\) is the average length of documents for a given document collection.

3.5 ATC weighting scheme

ATC [6,7] stands for Augmented TFIDF weighting scheme. It is given by:

\[
  w_{ij} = \left( 0.5 + 0.5 \frac{x_{ij}}{max.x_{ij}} \right) \log \left( \frac{N}{n_j} \right) \sqrt{\sum_{i=1}^{N} \left[ \left( 0.5 + 0.5 \frac{x_{ij}}{max.x_{ij}} \right) \log \left( \frac{N}{n_j} \right) \right]^2}
\]

where \(x_{ij}\) is the term frequency, \(max.x_{ij}\) is the maximum term frequency in each document, \(N\) is the total number of documents in the collection, \(n_j\) is the number of documents in which term \(j\) occurs.

3.6 TFICF weighting scheme

The TFICF [7] refers to term frequency-inverse corpus frequency. This scheme produces document clusters that are of comparable quality as those generated by other weighting schemes. TFICF weighting scheme is given by,

\[
  w_{ij} = \log (1 + x_{ij}) \log \left( \frac{N + 1}{n_j + 1} \right)
\]
where $N$ is the total number of documents in the corpus, $n_j$ is the number of documents in the corpus where term $j$ occurs after removing the stopwords and applying Potter’s algorithm, $x_{ij}$ is the term frequency.

Among the weighting schemes, TFIDF is the most popular general-purpose term weighting scheme followed by Okapi.

### 4 Experimental Results

Our interest in this paper lies in studying the impact of the 6 term weighting schemes. We use the various weighting schemes mentioned in section 3 to cluster the documents by applying the $k$-means algorithm. For our experimental analysis, we consider a corpus containing 6 documents and also a benchmark data set `classic`. The `classic` data set contains 7095 documents classified into 4 classes.

#### 4.1 Comparison of Cluster Quality Measures for 6 documents

We compared the clustering solutions obtained for the different weighting schemes using cluster quality measures like entropy and purity. These are standard measures that help us to ascertain the cluster quality. The entropy of a good cluster should be low and that of purity should be high. Given below are the entropy and purity values of the clustering solutions obtained for the 6 term weighting schemes.

<table>
<thead>
<tr>
<th></th>
<th>TF</th>
<th>TFIDF</th>
<th>MI</th>
<th>ATC</th>
<th>Okapi</th>
<th>TFICF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Entropy</td>
<td>0.6667</td>
<td>1.0986</td>
<td>0.8675</td>
<td>1.0114</td>
<td>1.0986</td>
<td>1.0986</td>
</tr>
<tr>
<td>Purity</td>
<td>0.6667</td>
<td>0.6667</td>
<td>0.6667</td>
<td>0.6667</td>
<td>0.6667</td>
<td>0.6667</td>
</tr>
</tbody>
</table>

Table 1: Cluster Quality for 6 documents

From Table 1, we see that the purity values are the same for all term weighting schemes. Comparing the clustering solutions based on entropy, we see that the clustering solution obtained using the TF weighting scheme is better. It has the lowest entropy value.
4.2 Comparison of Cluster Quality for classic data set

In this section we present the comparison of the cluster quality of the clustering solutions for the classic data set for the six different weighting schemes. The k-means algorithm is applied to cluster the 7095 documents in this data set into 4 clusters.

<table>
<thead>
<tr>
<th></th>
<th>TF</th>
<th>TFIDF</th>
<th>MI</th>
<th>ATC</th>
<th>Okapi</th>
<th>TFICF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Entropy</td>
<td>0.7388</td>
<td>0.65965</td>
<td>0.9086</td>
<td>0.8051</td>
<td>0.47635</td>
<td>0.4432</td>
</tr>
<tr>
<td>Purity</td>
<td>0.4954</td>
<td>0.61240</td>
<td>0.4516</td>
<td>0.6214</td>
<td>0.7113</td>
<td>0.7620</td>
</tr>
</tbody>
</table>

Table 2: Cluster Quality for classic data set

From Table 2, we see that TFICF has the highest purity value and lowest entropy value which makes it clear that for classic data set TFICF weighting scheme gives the best clustering solution.

5 Conclusion

The impact of term weighting schemes on Informational Retrieval models has been well studied. But not much work has been done on the impact of term weighting schemes on text clustering. A good term weighting scheme will be able to give an informative term document matrix representation for the documents. Conversely, a bad term weighting scheme will only be able to provide small amount of content information about the documents.

From our analysis we notice that we cannot identify the supremacy of a particular term weighting scheme for the small document corpus that we have considered. But using a large data set we are able to identify which one is better. Also, the existing term weighting schemes do not distinguish between informative words from the un-informative ones, which is crucial to the performance of document clustering.

The role of term weighting schemes in text clustering needs to be further explored for different sizes and types of document collections.
References


