Automatic categorization of Journal Submission System: An approach based on Cosine Similarity

Wazhmma Ahmadi  
a.wazhmma@yahoo.com  
Department of Studies in Computer Science,  
University of Mysore, Karnataka

Dr. Hanumanthappa J.  
hanumbe@gmail.com  
Department of Studies in Computer Science,  
University of Mysore, Karnataka

ABSTRACT

In this paper, we have designed the automatic categorization of electronic journal submission system. Hence, eliminating the manual identification/selection of category and subsequently uploading the paper under that category. The proposed system reads the content, process it, removes the stop words and compares the remaining words with already set aside words for categorization/classification. From the extensive experimentation, it was revealed that the proposed system is robust and eliminates the manual involvement of submission administrator.

Keywords — Categorization/Classification, Submission System, Stop Words, Administrator

1. INTRODUCTION

With the rapid increase in the research and development, we see exponential increase in the submission and publishing of electronic journals/articles. After writing a research paper, a researcher uploads the paper to his chosen/selected journal electronic submission system. Thereafter, at first, system administrator has to manually read the paper, next decide the category of subject and finally assign the research paper to a particular submission system. Thus, involves large human effort, therefore, we need a system that is capable of performing these tasks automatically.

We explored the text/data mining concepts for our proposed system, for two reasons. First, it provides different prepossessing steps to transform unstructured data to structured data (information). Second, it extracts various types of information from textual sources. Broadly speaking, text mining falls in the broad category of natural language processing (NLP) that facilitates the communication (or interfaces) between human and computer. In literature, there exists various types of studies that have explored the text mining for various classifications. In [1], after following the steps involved in text mining, they were successful in classifying the news. In various studies, different machine learning techniques were explored for the classification of textual information [2,3,4,5]. Like, in [3], supervised classification technique is used for the categorization. Nevertheless, it still remains a challenge to get effective categorization of textual data, forget about automating journal electronic submission system. The rest of the paper is divided into following. Proposed system is described in Section 2, experimentation is carried out in section 3, and finally conclusion and future work is drawn in section 4.

2. PROPOSED SYSTEM

2.1 Pre-processing

In order to remove unnecessary characters/words, number and date formats, and other common words reoffered as stop words. This stage is an important stage, because output from this stage are fundamental units which are later used for categorization/classification.

2.1.1. Tokenization: It refers the breaking/dividing of steam of text into meaningful elements, like, characters, words, and symbols generally referred as tokens.

2.1.2. Stop word removal: Stop words are used to join words in a sentence. It has been observed thee do not contribute context (or meaning) textual content. Furthermore, they exist more frequently in the data, hence, it is better to remove them at the outset.
2.1.3. Stemming: It gives common representation to different variant form of a word. For example words like, presentation, presented, and presenting can be reduced to single/common word present. In other words, reducing words to their root word [6,7]. The main motivation for this step is to reduce data and get meaningful words from huge amount of words present in the text.

\[ TF(t) = \frac{\text{Frequency of term } t \text{' in a given document}}{\text{Total terms in a document}} \]

\[ IDF(t) = \log \frac{\text{Total number of documents}}{\text{Total number of documents which contain term } t} \]

2.3 Cosine similarity

It measures the similarity between the documents without bothering about their size. After measuring the cosine of an angle between the two vectors (word counts of documents) it projects the angle in a multi-dimensional space [6,7,8]. Formally,

\[ \text{Similarity} = \cos(q) = \frac{A \cdot B}{\|A\| \|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}} \]  

In equation 1, the terms \( A_i \) and \( B_i \) are components of vector A and B, respectively. We get following ranges of the results:

- 1: Opposite, 1: exactly same, and [1,1]: intermediate similarity

For example, we have a three documents that contain information about leading cricketers, (i) Ricky Pointing (RP), (ii) Hashim Amla (HA), and (iii) Sachin Tendulkar (ST), see Figure 2.

Let the three words (taken arbitrarily) from these documents have following frequency (shown in Table 1).

<table>
<thead>
<tr>
<th>Terms</th>
<th>RP doc</th>
<th>HA doc</th>
<th>ST doc</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clean Bold</td>
<td>10</td>
<td>50</td>
<td>200</td>
</tr>
<tr>
<td>Hit Wicket</td>
<td>400</td>
<td>100</td>
<td>20</td>
</tr>
<tr>
<td>Catches Taken</td>
<td>10</td>
<td>5</td>
<td>1</td>
</tr>
</tbody>
</table>
Table 2: Similarity matrix based on cosine similarity (see equation 1)

<table>
<thead>
<tr>
<th>Similarity score</th>
<th>Sum of common words</th>
<th>Cosine similarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>RP doc and HA doc</td>
<td>10 + 50 + 10 = 70</td>
<td>0.15</td>
</tr>
<tr>
<td>HA doc and ST doc</td>
<td>20 + 10 + 7 = 37</td>
<td>0.23</td>
</tr>
<tr>
<td>ST doc and RP doc</td>
<td>10 + 10 + 7 = 27</td>
<td>0.77</td>
</tr>
</tbody>
</table>

From Table 3 as one can observe all three documents are connected with a cricket. Therefore, in cosine similarity objective is the quantitatively measure the similarity between given documents.

2.4. Labelling
Once the cosine similarity approach finds the similarities between a testing document and predefined categories (training set), the system searches for the highest similarity and label it to that category.

3. EXPERIMENTATION
We have created dataset which contains following categories Arts, Economics, Political, and Sports. We divided the dataset into 75:25 and 80:20 of training and testing sets respectively. After following each step (see section 2) we trained our proposed system after computing cosine similarity measure and labelling (see Section 2.3 and section 2.4). Later, proposed system is evaluated on independent test sets. Note, we evaluated the performance of the proposed system in $F$- measure [9]. Further, experiments are repeated for 20 random trials, later average $F$- measure is computed. Table 4 presents the overall experimental results for both of the ratios.

Table 3: Recognition rate (in terms of F-measure) obtained for different ratios of the dataset.

<table>
<thead>
<tr>
<th>Ratio</th>
<th>Arts</th>
<th>Economics</th>
<th>Political Science</th>
<th>Sports</th>
<th>Overall recognition rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>75:25</td>
<td>0.75 ± .01</td>
<td>0.78 ± .001</td>
<td>0.79 ± .001</td>
<td>0.80 ± .010</td>
<td>0.79 ± .110</td>
</tr>
<tr>
<td>80:20</td>
<td>0.76 ± .01</td>
<td>0.79 ± .001</td>
<td>0.80 ± .001</td>
<td>0.82 ± .010</td>
<td>0.81 ± .001</td>
</tr>
</tbody>
</table>

From Table 4 as number of training samples are increased, we observe increase in recognition performance.

4. CONCLUSION AND FUTURE WORK
To sum up, the paper focuses on categorizing and organizing articles automatically by the system which enhances the efficiency of retrieval and publishing the articles and journals by users. It reads the content, processes it and classifies/categories the articles based on training documents. The efficiency and accuracy of the system are determined by usage of appropriate preprocessing, representation, similarity approach used as well as the dataset in training phase. To have a high-performance system and increase the scope of the system, a couple of works have to be done. Such as increasing the corpus of the dataset and observe the efficacy of the proposed system. Second, making the system able to create new categories while a document that does not belong to either of the predefined categories.

5. REFERENCES