Introduction:

Similarity measures have been extensively used in text classification and clustering algorithms. Several characteristics are embedded in this measure. It is a symmetric measure. The difference between presence and absence of a feature is considered more essential than the difference between the values associated with a present feature. The similarity increases as the difference between the two values associated with a present feature decreases. Furthermore, the contribution of the difference is normally scaled. The similarity decreases when the number of presence-absence features increases. An absent feature has no contribution to the similarity.

The proposed measure is extended to gauge the similarity between two sets of documents. The measure is applied in several text applications, including single-label classification, multi-label classification, k-means like clustering, and hierarchical agglomerative clustering, and the results that will be obtained will demonstrate the effectiveness of the proposed similarity measure.
Objectives of the project:

The main objective is to compute the similarity between the two documents with respect to texts or images. First case similarity increases as the difference between two involved feature values decreases. Second case a fixed value is contributed to the similarity. The similarity increases as the difference between the two values associated with a present feature decreases. The similarity decreases when the number of presence-absence features increases. An absent feature has no contribution for the similarity.

Methodology:

Module Flow Diagram:

Keyword Extractor Module

Start

Scan the document for words

Display distinct words to the user

Stop
Keyword Score Generator Module

Start

Load keywords

Get the documents

Calculate the count of the keywords

End

Keyword Score Generator Module

SMTP Comparison Module

Start

Get the score of reference

Get the scores of other document

SMTP comparison process

Apply SMTP formula

Generate comparison result

Stop

SMTP Comparison Module
SMTP formula:
\[ S_{SMTP}(G1, G2) = F(G1, G2) + \lambda \]
\[ 1 + \lambda \]

Where G1 and G2 are the two document sets
\( \lambda \) is the user constant

Image Comparison

Start

Fetch images

Rescale & partition image

Comparison process

If difference

No

Discard image

Yes

Verification process

Convert into original image

Display result to user

End
Data Flow Diagrams

A data flow diagram (DFD) is a graphical representation of the "flow" of data through an information system, modeling its process aspects. Often they are a preliminary step used to create an overview of the system which can later be elaborated. DFDs can also be used for the visualization of data processing (structured design).

DFDs help system designers and others during initial analysis stages visualize a current system or one that may be necessary to meet new requirements. Systems analysts prefer working with DFDs, particularly when they require a clear understanding of the boundary between existing systems and postulated systems. DFDs represent the following:
1. External devices sending and receiving data
2. Processes that change that data
3. Data flows themselves

Keyword Extractor Module

Reads the documents to be compared for the words performs search on each word and fetches the unique words from the document, integrates it and generates the list with unique keywords.
Keyword Score Generator Module

Calculates the count of each word from the document and provides the list of keywords and their count.

SMTP Comparison Module

Searches in the document for the keyword in two documents and get the similarity details of the two keywords being searched.
In system requirements we used Java as language for coding and Netbeans as a platform, with minimum basic requirements of a system. In system design part we represented a system architecture that discussed how a client connects with a browser to retrieve texts and images stored in various servers. In the methodology we discussed two different techniques

- Flow Diagram- which describes each modules and their interconnections.
- Data Flow Diagram- which describes the different data’s where those data’s are initialized and where they are used.
Results:

Similarity Measure

Copyright © All Rights Reserved.
Unique Keywords of Other Documents:

- apple, banana, mango
- jackfruit, mango, apple
- apple, banana

Count Feature:

Count Feature of Reference Document:

- apple: 1
- banana: 1
- mango: 1
- jackfruit: 1

Count Feature of Other Documents:

- apple: 1
- banana: 1
- mango: 1
- jackfruit: 1
Similarity Measure

<table>
<thead>
<tr>
<th>Reference Document</th>
<th>Other Document</th>
<th>Similarity Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>C:/Users/RAHESH/Desktop/images/Similarity Measure build web uploaded/A.txt</td>
<td>C:/Users/RAHESH/Desktop/images/Similarity Measure build web uploaded/A.txt</td>
<td>0.0</td>
</tr>
<tr>
<td>C:/Users/RAHESH/Desktop/images/Similarity Measure build web uploaded/A.txt</td>
<td>C:/Users/RAHESH/Desktop/images/Similarity Measure build web uploaded/B.txt</td>
<td>0.425</td>
</tr>
<tr>
<td>C:/Users/RAHESH/Desktop/images/Similarity Measure build web uploaded/A.txt</td>
<td>C:/Users/RAHESH/Desktop/images/Similarity Measure build web uploaded/C.txt</td>
<td>0.0</td>
</tr>
<tr>
<td>C:/Users/RAHESH/Desktop/images/Similarity Measure build web uploaded/A.txt</td>
<td>C:/Users/RAHESH/Desktop/images/Similarity Measure build web uploaded/D.txt</td>
<td>0.0</td>
</tr>
<tr>
<td>C:/Users/RAHESH/Desktop/images/Similarity Measure build web uploaded/A.txt</td>
<td>C:/Users/RAHESH/Desktop/images/Similarity Measure build web uploaded/E.txt</td>
<td>0.0</td>
</tr>
</tbody>
</table>
Conclusions and Future Work:

We have presented a novel similarity measure between two documents. Several desirable properties are embedded in this research area. However, it would be of greater value evaluating the performance of the measures on larger test-beds. Also, this work mainly focuses on textural features. It would be interesting to investigate the effectiveness and efficacy of our proposed model in the scenarios that involve non-textual features and objects. Besides, as can be seen from the experimental results, the usefulness of a similarity measure could depend on (1) application domains, e.g., text or image, (2) feature formats, e.g., word count or tf-idf, and (3) classification/clustering algorithms. It would be a very interesting topic to examine how certain similarity measures behave in different classification/clustering tasks.