PARADIGMS OF PROGRAMMING

MINI-PROJECT 2

Expert Systems

OBJECTIVE

Building an automated system which matches Reviewers(Experts) and Candidates of any conference.

ALGORITHM (v 1.0)

A content based filtering algorithm has been used.

Reviewer side Algorithm :

Each reviewer has a corpus of documents which contains abstracts of his previously published papers.

TF-IDF algorithm is applied over this corpus of documents which gives us a set of words for each document with a score ,then we take mean of all these vectors(contains words with a score) and get a final vector from this. From this vector 20% of the words are selected which have the highest weight . These words represent the reviewer profile.

Candidate side Algorithm:

Now the paper to which a reviewer is to be assigned is selected and Jaccard similarity function is applied between the set of words of the reviewer as calculated above and the set of words of the paper. By this process we assign score to each reviewer on a research paper.

FLOW CHART

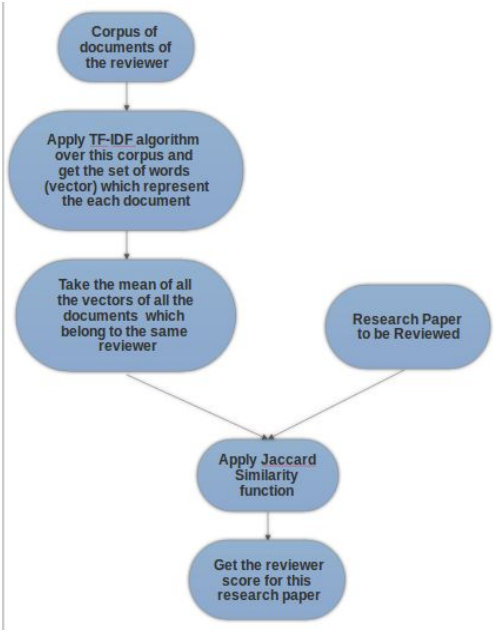

FIG. v 1.0

ALGORITHM (v 1.1)

Reviewer side Algorithm :

Initially we took the mean of all the vectors(set of words in each document), now for a particular word we take the highest score from all the vectors (each vector is the output from the TF-IDF for each document) and assign this highest score to this particular word and form the final vector. Now the final vector represent the reviewer profile.

Candidate side Algorithm:

Same as algorithm v 1.0.

FLOW CHART

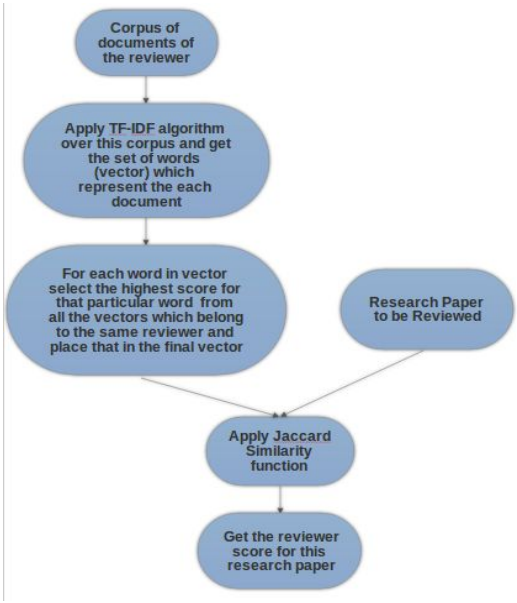

FIG. v 1.1

ALGORITHM (v 1.2)

Reviewer side Algorithm :

K-Means algorithm is applied over all documents of all the reviewers which gives us clusters of documents .

TF-IDF algorithm is applied over the documents present in a cluster and the output is a vector for each document in the cluster which contains a set of words that represent the document .

Mean of all the vectors of the documents which correspond to same reviewer is taken . This vector represents the reviewer profile.

Candidate side Algorithm:

Same as algorithm v 1.0.

FLOW CHART

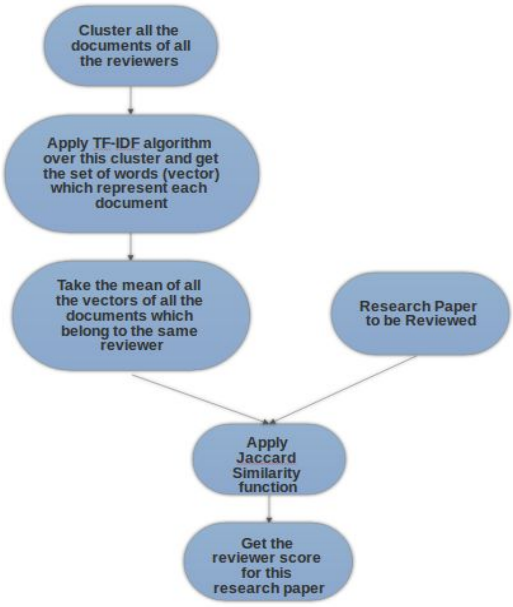

FIG. v 1.2

DATA SETS

1. http://turing.iitpkd.ac.in/moodle/pluginfile.php/5883/mod_assign/intro/Expert_Finding_Data.zip
2. http://turing.iitpkd.ac.in/moodle/pluginfile.php/5883/mod_assign/intro/Reviewer_Info.zip
3. http://turing.iitpkd.ac.in/moodle/pluginfile.php/5883/mod_assign/intro/title_abstract_data.tar

DATA STRUCTURES

1. Dictionary
2. List
3. set

RESULTS

The output.txt file contains the reviewer scores for each research paper

CONTRIBUTIONS

PRABAL VASHISHT

Design of TF-IDF algorithm (clustering) ,Parsing the reviewer data,research,report

RAJAT SHARMA

Design of Jaccard similarity function, GUI design, research ,report

ROHITH REDDY G

Design of TF-IDF algorithm (max,mean) ,Parsing the candidate data,research ,report

REFERENCES

1. The Toronto paper matching system: an automated paper-reviewer assignment system L Charlin, R Zemel - 2013
2. Wikipedia
3. scikit-learn.org

APPENDIX

TF-IDF algorithm

Term frequency–inverse document frequency is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus

Term frequency

In the case of the **term frequency** tf(*t*,*d*), the simplest choice is to use the *raw count* of a term in a document, i.e. the number of times that term *t* occurs in document *d*. If we denote the raw count by $f_{t,d}$, then the simplest tf scheme is tf(*t*,*d*) = $f_{t,d}$

$$\mathrm{tf}(t,d) = 0.5 + 0.5 \cdot \frac{f_{t,d}}{\max\{f_{t',d} : t' \in d\}}$$

Inverse document frequency

The ***inverse document frequency*** is a measure of how much information the word provides, that is, whether the term is common or rare across all documents. It is the logarithmically scaled inverse fraction of the documents that contain the word, obtained by dividing the total number of documents by the number of documents containing the term, and then taking the logarithm of that quotient.

$$\mathrm{idf}(t, D) = \log \frac{N}{|\{d \in D : t \in d\}|}$$

N : Total no of documents

D : No of documents which contain the term t

Jaccard Index

The **Jaccard index**, also known as **Intersection over Union** and the **Jaccard similarity coefficient** is a statistic used for comparing the similarity and diversity of sample sets. The Jaccard coefficient measures similarity between finite sample sets, and is defined as the size of the intersection divided by the size of the union of the sample sets:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}.$$