Community Detection for testing hypothesis of dispersion similarity

Anudeep PV and Vellanki Sai Harsha

Abstract— **A scientific domain consists of subfields that can be further refined into specialisations. Specialisations emerge, evolve and consolidate, as reflected in particular in literature development, along a contents-based dimension where important problems are stated and addressed,and along a communal dimension where researchers collaborate and compete to solve those problems. We propose a generic framework that aims at effectively identifying and characterising the main specialisations of the subfields of a scientific domain by leveraging paper contents.More specifically, the latent knowledge structure of a domain is discovered and progressively refined along both the contents-based and communal dimensions**

I. INTRODUCTION

Take a specific domain like natural language processing in computer science field.It consists of a number of sub-fields : Lexical Analysis, Speech Recognition, Machine Translation, etc. These subfields are themselves subdivided into specialisations.For instance, Lexical Analysis encompasses the specialisations of Word Boundary Detection, Grammatical Category Assignment. . . ; Speech Recognition encompasses the specialisations of Acoustic Signal Processing, Temporal Pattern Mining. . . ; Machine Translation encompasses the specialisations of English to Chinese Machine Translation, English to French Machine Translation. . .

Specialisation consists of a number of closely related studies that jointly endeavour to solve particular problems or develop certain techniques, based on the efforts of a community of researchers who closely work together, whether they collaborate or compete, whether they use each others work or oppose their frameworks or experiments.Therefore, a specialisation comes to existence and develops along two dimensions: a contents-based dimension along which we find knowledge, problems, methods, etc.; a communal dimension along which we find citations, work-shops, etc.

In this project we mainly concentrated to divide research papers into some sets such that each set consisted of closely related research papers.In Section II ,we describe what is the dataset we took and how did we pre-process the data.In section III,we describe what algorithms used and brief explanation of that algorithm.In section IV, we describe the results we got by running that algorithm on given dataset by changing some parameters in algorithm.

II. PRE-PROCESSING OF DATA

A. Dataset

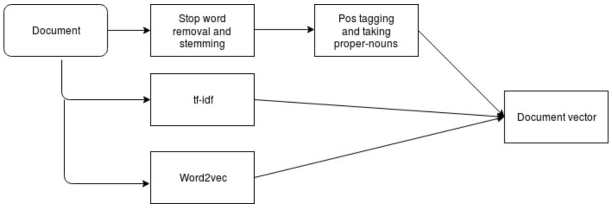
The dataset considered here is set of 346 research papers from various domains in computer science field.

B. Pre-Processing of that data

We used pdf2txt to convert given set of research papers to text files,then we removed all the stop words from data and then did stemming on the data.We then used word2vec library to convert given word to a vector.Then we assigned weights to each word using tf-idf method.For a given document we applied pos tagging and then took only proper-nouns(NNP) as key words. For these words we used tf-idf as weight and multiplied it with corresponding word-vector.This is taken as document vector.

For each document,we calculate document vector as follows:

$$Doc - vector = \sum_{for\ each\ NNP} tf - idf * word - vec$$


Document
Stop word removal and stemming
Pos tagging and taking proper-nouns
tf-idf
Word2vec
Document vector

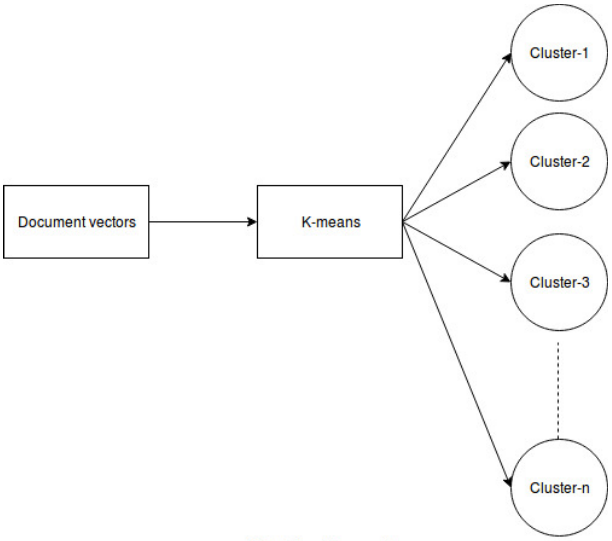
III. ALGORITHMS

We used k-means algorithm to cluster the data.We changed number of clusters and ran this algorithm on that dataset and observed the results.

k-means is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed apriori. The main idea is to define k centers, one for each cluster. These centers should be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other.

The next step is to take each point belonging to a given data set and associate it to the nearest center. When no point is pending, the first step is completed and an early group age is done. At this point we need to re-calculate k new centroids as barycenter of the clusters resulting from the previous step.

After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new center. A loop has been generated. As a result of this loop we may notice that the k centers change their location step by step until no more changes are done or in other words centers do not move any more. Finally, this algorithm aims at minimizing an objective function know as squared error function which is sum of squares of euclidian distances of a point from its cluster center.



Clustering Documents

IV. RESULTS

TABLE I
OBSERVATIONS

No of Clusters	Distribution of documents into clusters
3	116,121,109
5	66,54,56,89,81
7	50,32,71,29,54,37,73
9	30,28,45,72,27,41,32,55,16
11	32,21,46,25,61,26,29,10,30,41,25

V. CONCLUSIONS

We took odd number of clusters because k-means algorithm generally performs better when the number of clusters is odd and based on our results, we observed that if the number of clusters is 5/7,grouping is being done reasonably well.