Clustering and Visualization of University Webpages, CS 6140, Spring 2017

Avani Sharma (u1065441) and **Srivatsan Kumar (u1077227)**

School of Computing, University of Utah

APPROACH

Techniques Used:

- **Bag of Words:**
  Bag of Words model is a way to represent a text document such that the frequency of occurrence of each word except stop words (is, the, are, on etc.), is used as a feature to represent the document. Each document is expressed as vector, whose entries are the number of times each word occurs in that document.
- **Clustering:**
  Run multiple clustering algorithms such as K-means, Spectral, Ward, Birch and Agglomerative on the different sets of features and examine confusion matrices for each set.
- **Visualization:**
  Find the features with the most information gain. Represent the features in the dataset using a parallel co-ordinates representation with clusters and labels. Analyze the visualization through brushing.

Dataset

The Universities: Cornell, Texas, Washington, Wisconsin
The Labels: Student, Faculty, Project, Course

- **Dataset: 4 Universities Data Set** (January 1997), **8282** Web Pages, from **4** Universities, manually classified into 8 categories.
- **Used Preprocessed Dataset** , 4 Categories, Ana Cardoso Cachopo, "Subset of 4 Universities Dataset (Webkb)"
- **Data Preprocessing:**
  Convert the HTML documents to a set of space separated words and removed commonly used words from the document.
- **Feature Extraction:**
  1. Documents: as vectors containing the counts of the occurrences of words in the corpus.
  2. Extracted more relevant features : TF-IDF, PCA
- **Technologies:** *Python, ScikitLearn, Numpy, d3*

Did you know?
Different implementations of the same algorithm were found to exhibit enormous performance differences, with the fastest on a test data set finishing in 10 seconds, the slowest taking 25988 seconds.

INTRODUCTION

Problem Statement Given a large corpus of labelled mixed heterogeneous web pages, group similar web pages that represents a particular category and find distinct identifying characteristics for a particular category.

TF-IDF

$$\text{Term Frequency} = tf(t,d) = \frac{f(t,d)}{\sum_{t',d} f(t',d)} \quad (1)$$

where,
$f(t,d)$ = frequency of term t in document d
$\sum_{t',d} f(t',d)$ = Sum of frequencies of all the words in document

$$\text{Inverse Document Frequency} = idf(t,N) = \log \frac{N}{1+n_t} \quad (2)$$

where,
N = total number of documents
n_t = Number of documents where term t appears means
Adding 1 to denominator prevent division from 0

$$\text{TF-IDF} = tf(t,d) * idf(t,N) \quad (3)$$

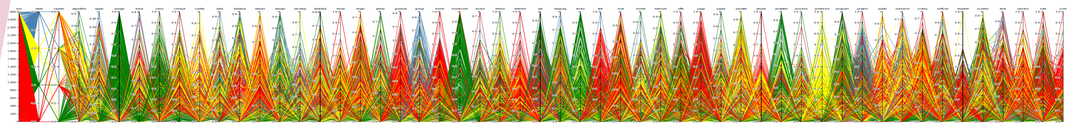

Fig. 1: Parallel co-ordinates representation of All Data points

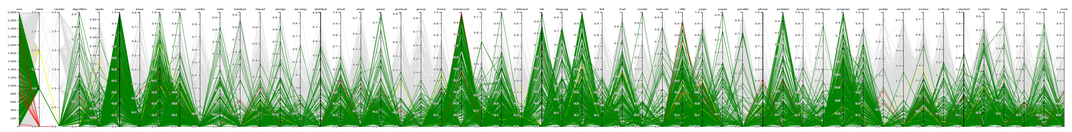

Fig. 2: Parallel co-ordinates representation of Cluster-0

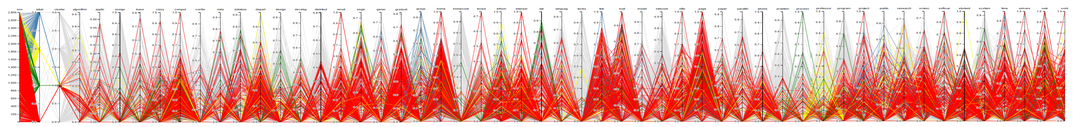

Fig. 3: Parallel co-ordinates representation of Cluster-1

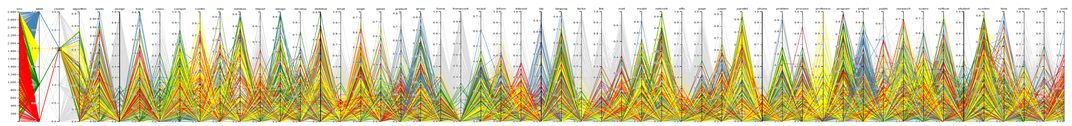

Fig. 4: Parallel co-ordinates representation of Cluster-2

RESULTS : CONFUSION MATRIX

On Features from Bag of Words

Labels → / Clusters ↓	Student	Course	Faculty	Project
0	1078	558	673	298
1	0	0	1	0
2	0	0	0	1
3	19	67	76	37

- Bag of Words Features perform poorly and are not representative of the dataset
- TF-IDF Features clearly separate labels, are better suited to the current task

On Features after vectorization using TF-IDF

Labels → / Clusters ↓	Student	Course	Faculty	Project
0	10	470	6	0
1	551	46	61	50
2	176	94	402	265
3	360	10	281	21

Did you know?
Parallel co-ordinates are also used in Air Traffic Control, Intrusion Detection and Optimization

VISUALIZATION

The figures are parallel co-ordinates representation of the clusters obtained from by running the K-means algorithm on the tf-idf features obtained. The colors represent the different labels: Student-Red, Course-Green, Faculty-Yellow and Project-Blue. The peaks in the figures represent words which have high tf-idf values

1. The first figure represents the entire dataset represented only by the tf-idf features with the most information gain.
2. Clusters 0 and 1 show clear separation between labels and their prominent features are the words with thick peaks
3. Cluster 2 contains a mix of different labels with no clear separations

Did you know?
The label Course is most associated with the words Assignment, Project, Program and Homework. What a surprise!

REFERENCES

1. Dell Zhang, Xi Chen, Wee Sun Lee. **"Text classification with kernels on the multinomial manifold"**
2. Thomas Finley, Thorsten Joachims, **"Supervised k-Means Clustering"**

Customized latex tikz poster, styles modified by **Avani Sharma** and **Srivatsan Kumar** LaTeX TikZposter