Text Classification Combining Clustering and Hierarchical Approaches

Shankar Ranganathan

MS Thesis Defense

May 3rd, 2004

Committee

Dr. Susan Gauch (Chair)

Dr. Perry Alexander

Dr. David Andrews

Department of Electrical Engineering and Computer Science

Presentation Outline

- Search Engines Today
- Contributions
- Related Work
- Text Classification – Our Approach
- Experiments and Evaluation
- Conclusions
- Future Work

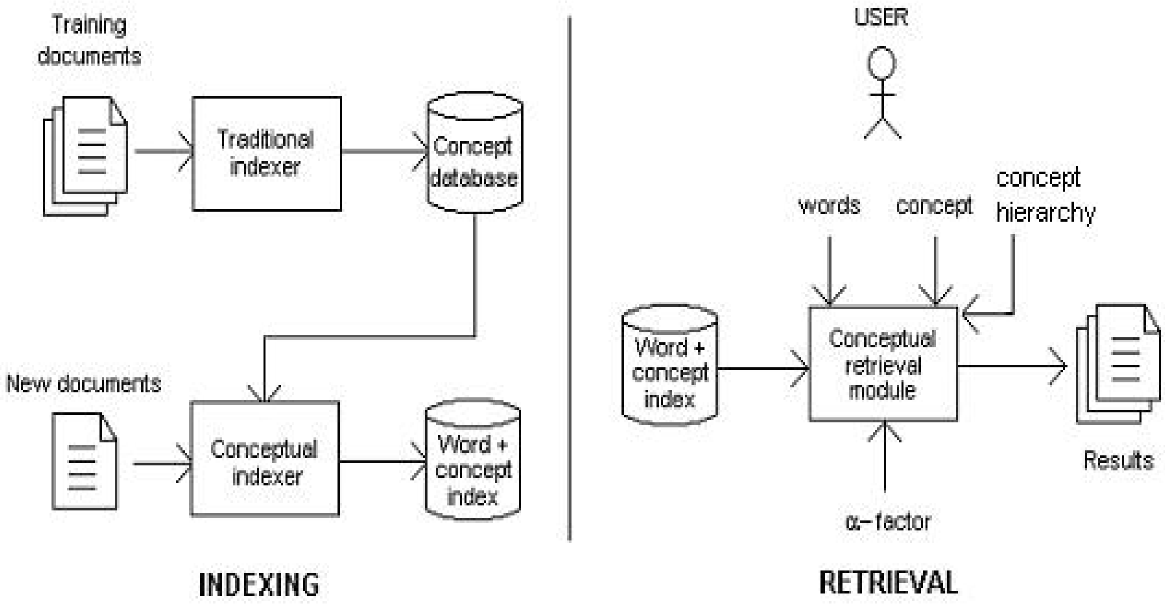
Search Engines Today

Return results based on simple key-word matches.

No regard for conceptual information.

For E.g. : If the query is “SALSA”, Is it……

KeyConcept Architecture



Contributions

- Novel approach to Text Classification by combining clustering within the concepts with hierarchical text classification
- Effect of clustering on flat classification versus hierarchical classification
- Effect of ignoring versus using concept wise distinction lower down the hierarchies

Related Work I

- **Text Classification**
  - Yang, Sebastiani: Comparison of Text classification methods - K-Nearest Neighbors, linear least square fit, Naïve Bayesian, Support Vector Machines, Decision trees
  - **Hierarchical Classification**: Proposed by Koller. Further work by – Sun, Labrou, Sasaki, Dumais, Wang

Related Work II

- Chaffee, YAHOO, Open Directory Project : **Ontology**
- Manning, Dubes, Kaufman – **Document clustering**

  Agglomerative (Guha, Karypis) vs. Divisive (Zhao)

  Lots of packages available on net – Cluto, Chameleon, Rock, Cure, DocCluster, Siftware etc.,
- Perkowitz – **Cluster Mining**

Text Classification

- Two Step Process : Training the classifier and Classification of new documents
- Training Phase:
  - Classifier is fed with documents that have been classified manually
  - Learns about the features (vocabulary) of the various categories into which new documents can be classified

Text Classification contd…

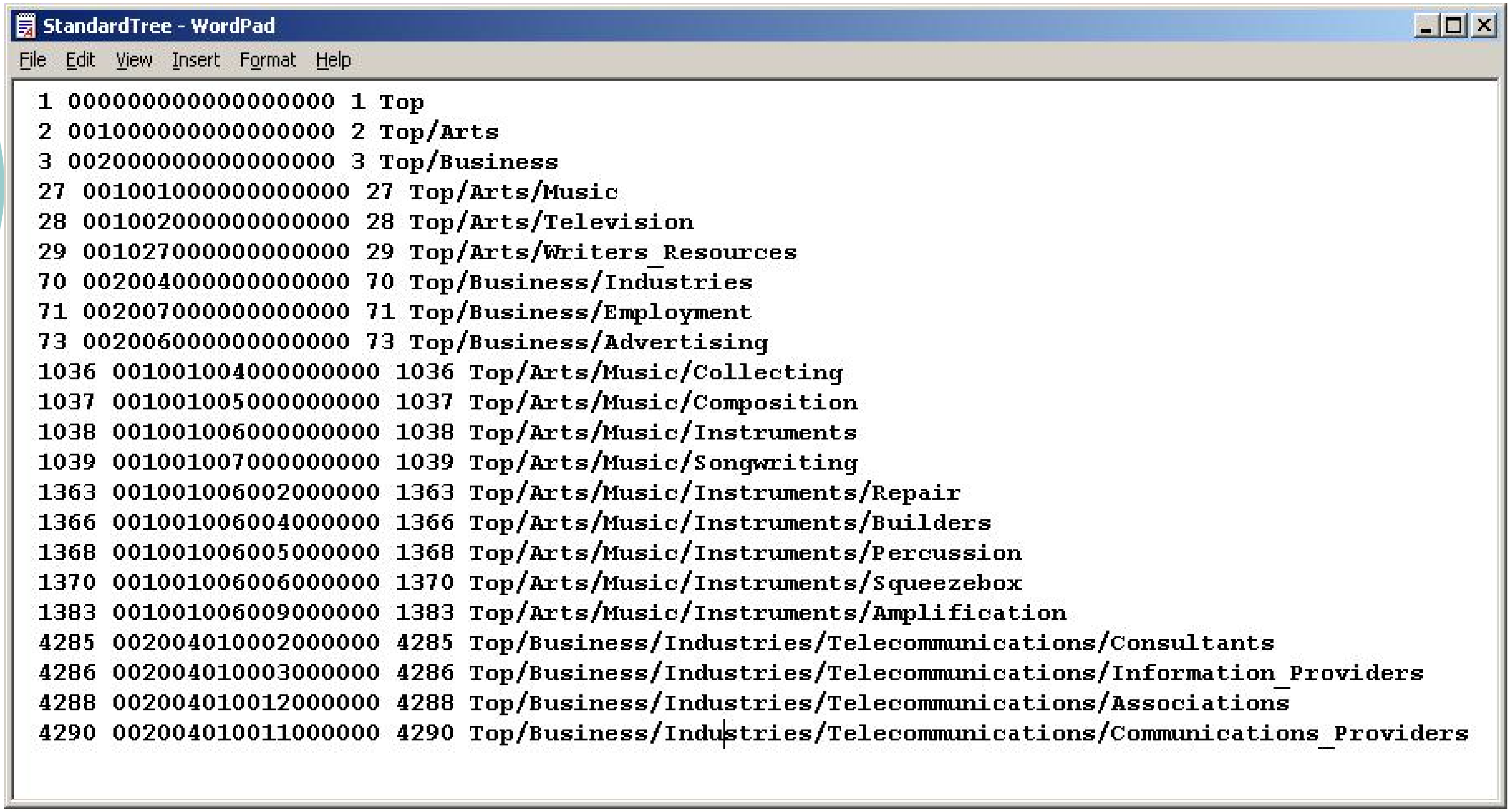
- Classification Phase:

  Classifier assigns category (ies) to new documents based on the similarity of the features of input document and of the categories that it learned during training

Text Classification – Our Approach

- Vector Space model (tf-idf)
- Training data are documents that are manually assigned to the categories Open Directory Project’s Standard Tree which is our reference Ontology
- Classifier creates a vector of vocabulary terms and associated weights in an inverted file

Standard Tree



```
1 000000000000000000 1 Top
2 001000000000000000 2 Top/Arts
3 002000000000000000 3 Top/Business
27 001001000000000000 27 Top/Arts/Music
28 001002000000000000 28 Top/Arts/Television
29 001027000000000000 29 Top/Arts/Writers_Resources
70 002004000000000000 70 Top/Business/Industries
71 002007000000000000 71 Top/Business/Employment
73 002006000000000000 73 Top/Business/Advertising
1036 001001004000000000 1036 Top/Arts/Music/Collecting
1037 001001005000000000 1037 Top/Arts/Music/Composition
1038 001001006000000000 1038 Top/Arts/Music/Instruments
1039 001001007000000000 1039 Top/Arts/Music/Songwriting
1363 001001006002000000 1363 Top/Arts/Music/Instruments/Repair
1366 001001006004000000 1366 Top/Arts/Music/Instruments/Builders
1368 001001006005000000 1368 Top/Arts/Music/Instruments/Percussion
1370 001001006006000000 1370 Top/Arts/Music/Instruments/Squeezebox
1383 001001006009000000 1383 Top/Arts/Music/Instruments/Amplification
4285 002004010002000000 4285 Top/Business/Industries/Telecommunications/Consultants
4286 002004010003000000 4286 Top/Business/Industries/Telecommunications/Information_Providers
4288 002004010012000000 4288 Top/Business/Industries/Telecommunications/Associations
4290 002004010011000000 4290 Top/Business/Industries/Telecommunications/Communications_Providers
```

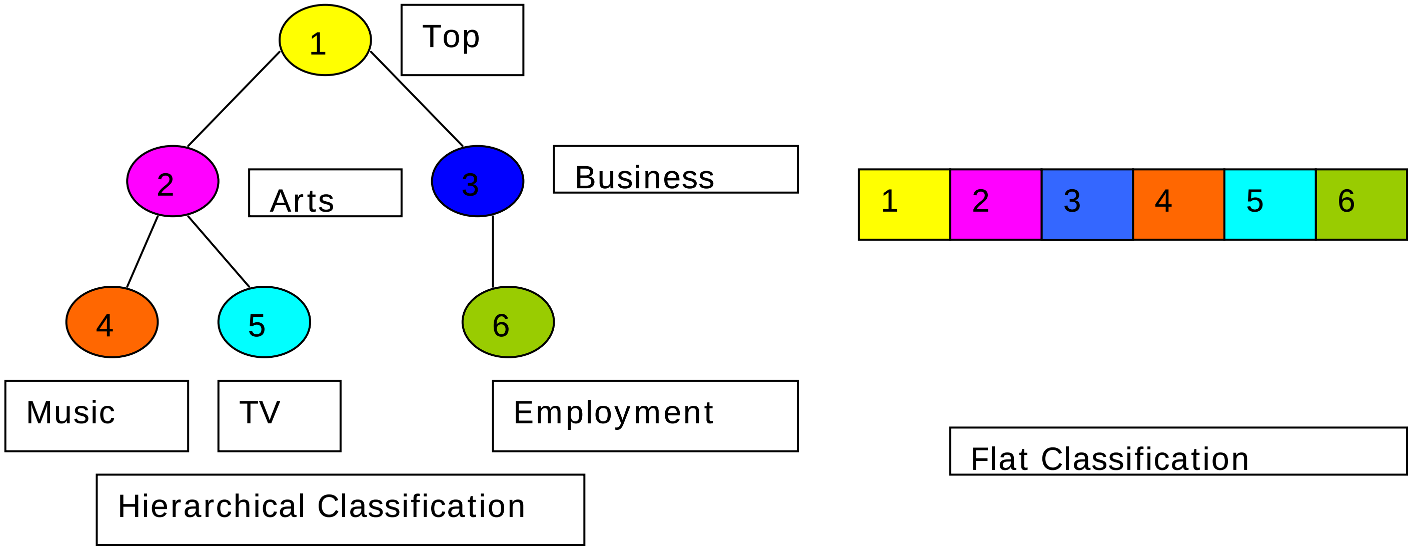
Text Classification – Our Approach ..

Feature selection during training (selecting training documents) plays a primary role towards improving classification accuracy.

○ Hierarchical classification

○ Use of Clustering

Flat Classification vs. Hierarchical Classification



- Top-down level-based Hierarchical classification

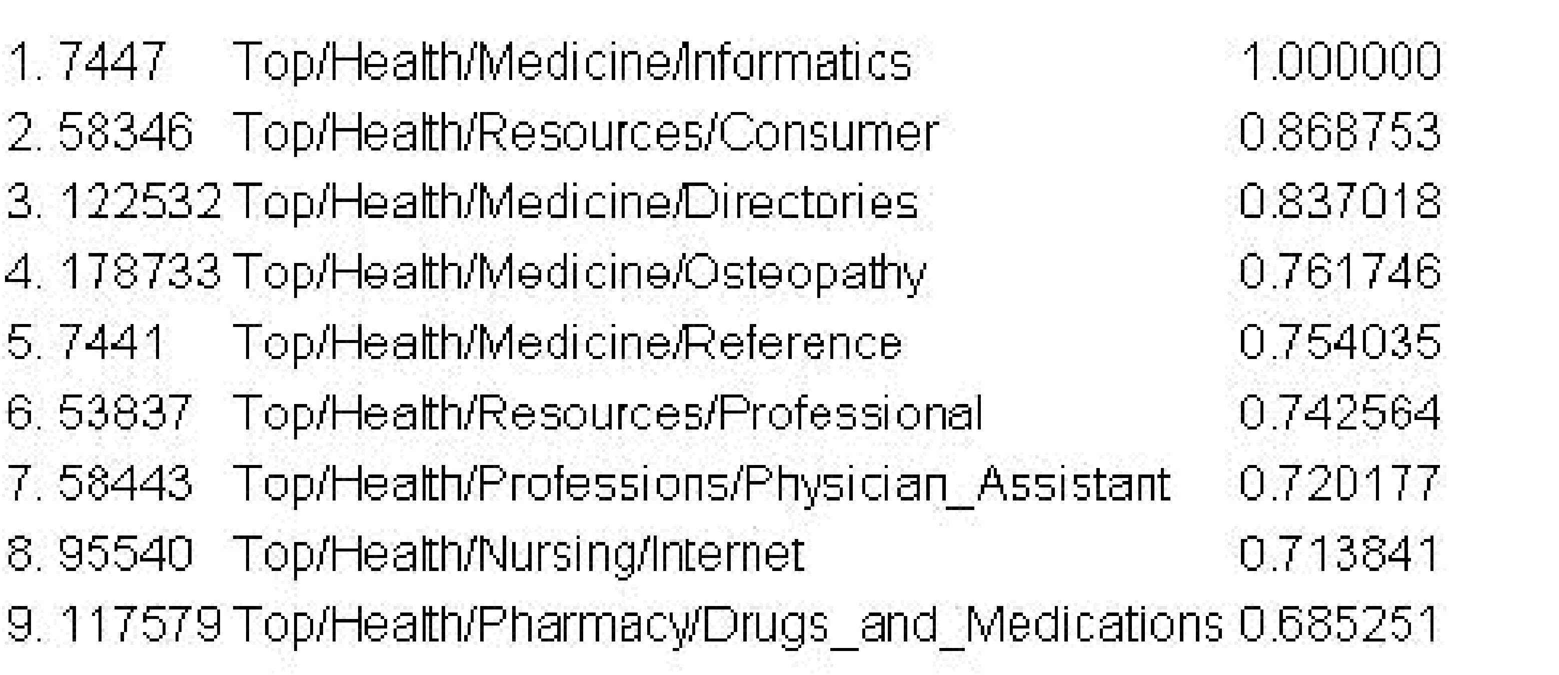
Role of Clustering

- Improve feature selection
- Eliminate documents that tend to confuse the classifier
- Identify within-category clusters, and extract cluster(s)' representative pages
- Document mining within the framework of cluster mining

Text Classification – Our Approach contd…

- During Classification phase, a vector of input document is created
- Similarity between training this vector and vector of each concept during training is computed using dot product
- New document is assigned to the categories with best matches

Classifier Output



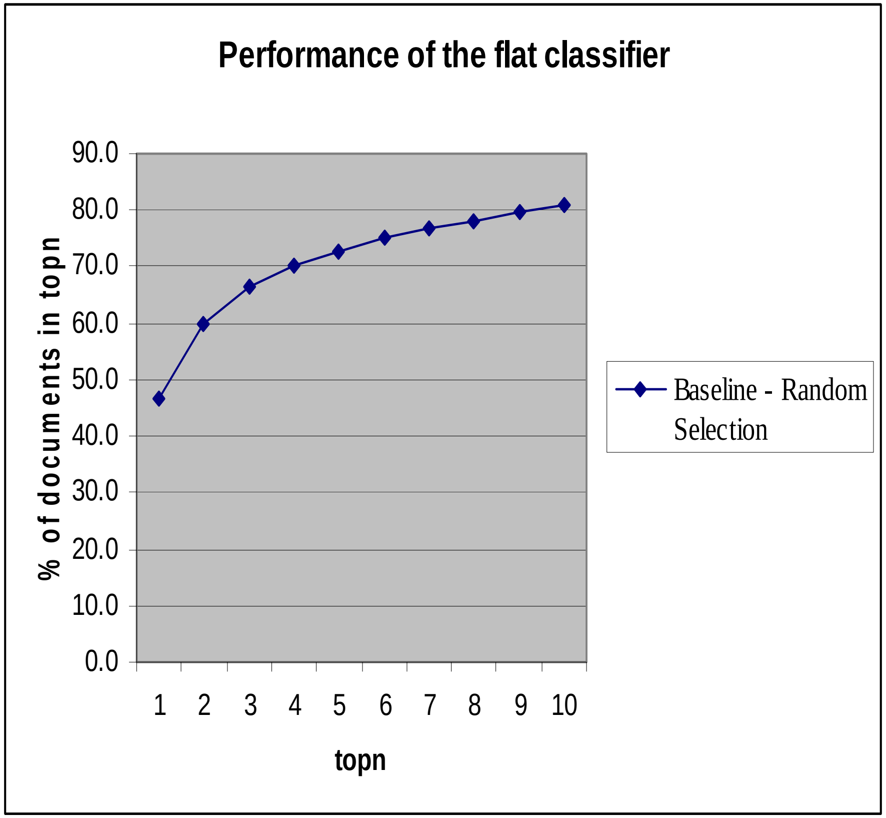
```
1. 7447    Top/Health/Medicine/Informatics                     1.000000
2. 58346   Top/Health/Resources/Consumer                       0.868753
3. 122532 Top/Health/Medicine/Directories                      0.837018
4. 178733 Top/Health/Medicine/Osteopathy                       0.761746
5. 7441    Top/Health/Medicine/Reference                       0.754035
6. 53837   Top/Health/Resources/Professional                   0.742564
7. 58443   Top/Health/Professions/Physician_Assistant          0.720177
8. 95540   Top/Health/Nursing/Internet                         0.713841
9. 117579 Top/Health/Pharmacy/Drugs_and_Medications 0.685251
```

Experimental Set-up

- Source of training data: Open Directory Project (dmoz.org) – ODP ontology contains hierarchical information
- Test data: Randomly-selected level 3 documents
- Clustering package: CLUTO
  - Clustering method: Partitional clustering
  - Similarity function: Cosine function
  - Program used: *vcluster - zscores*

Experimental Setup…..
Baseline – Random Selection

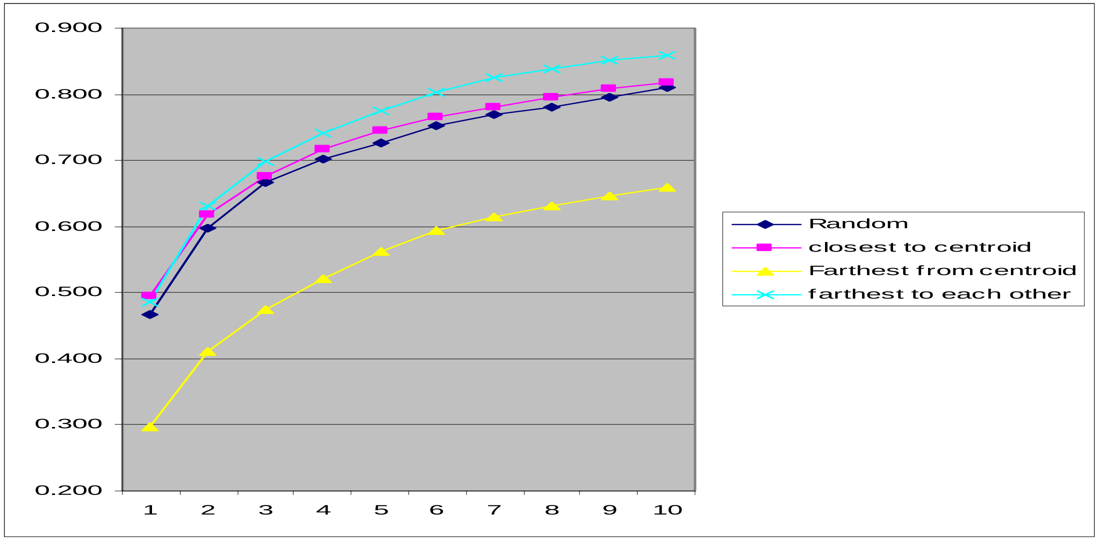
- All concepts from levels 1, 2 and 3 with at least 32 documents (total 1484)
- 2 documents from each concept was randomly withheld for testing (total - 2978)
- Trained with randomly-selected 30 documents from each concept(around 44500)
- Accuracy = 46.6 %



Evaluation

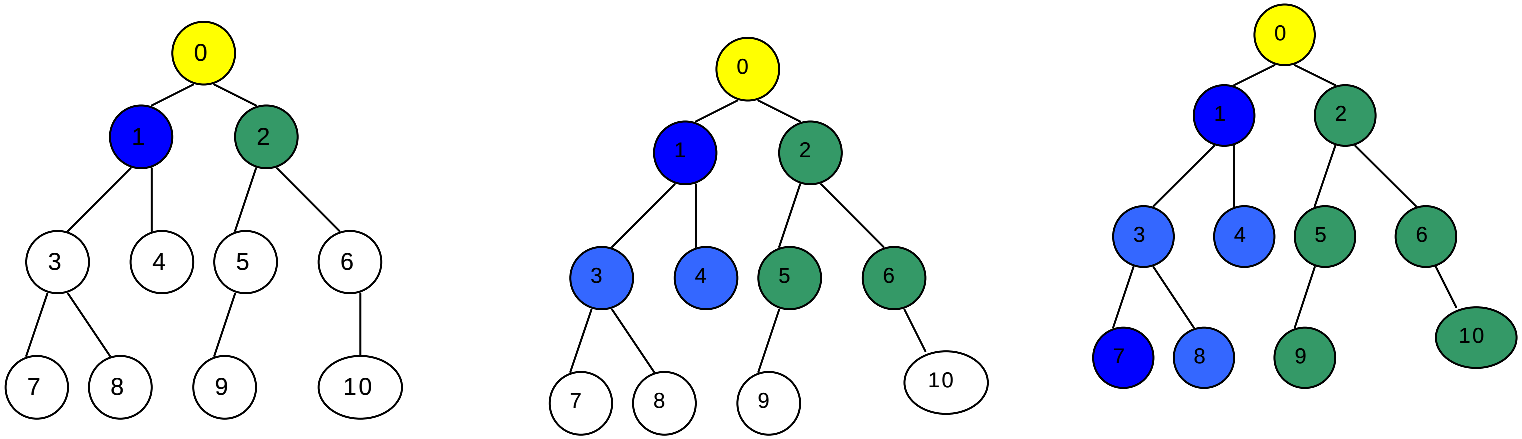
- Does selecting documents closest to the centroid to train improve accuracy ?
- For hierarchical classification, how far down the hierarchy should we go in each step ?
- What is the number of documents to train the classifier to get best results ?
- ‘Ignore’ or ‘consider’ tree structure among children ?

Experiment 1 : Effect of clustering on Flat Classification



- Best observed accuracy – Selecting documents closest to the centroid (49.5%)
- Poor performance – Selecting documents farthest from the centroid (29.5%)
- Selecting documents farthest from each other – 48.6%

Experiment 2- Effect of clustering on training Set selection for hierarchical classification

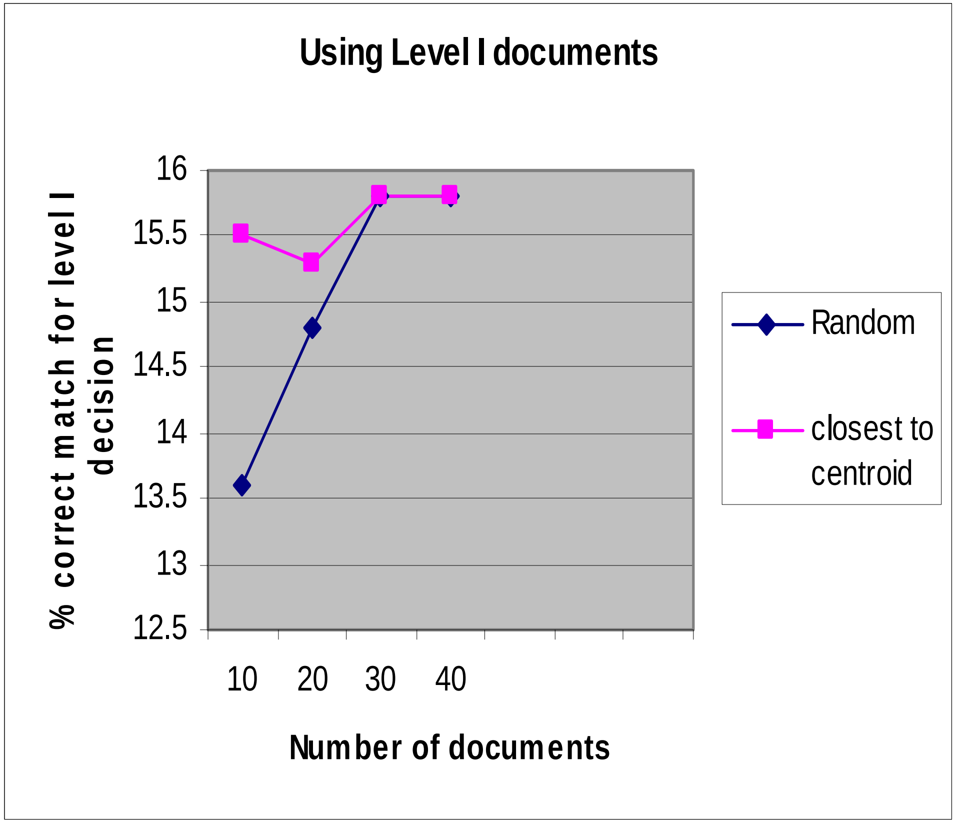


- 1 Classifier at level 1, 15 at level 2, 358 at level 3
- Documents from parent & children (& grandchildren put in the same pool to select)
- Parameters we tune : Depth, Random selection vs. clustering, # of documents

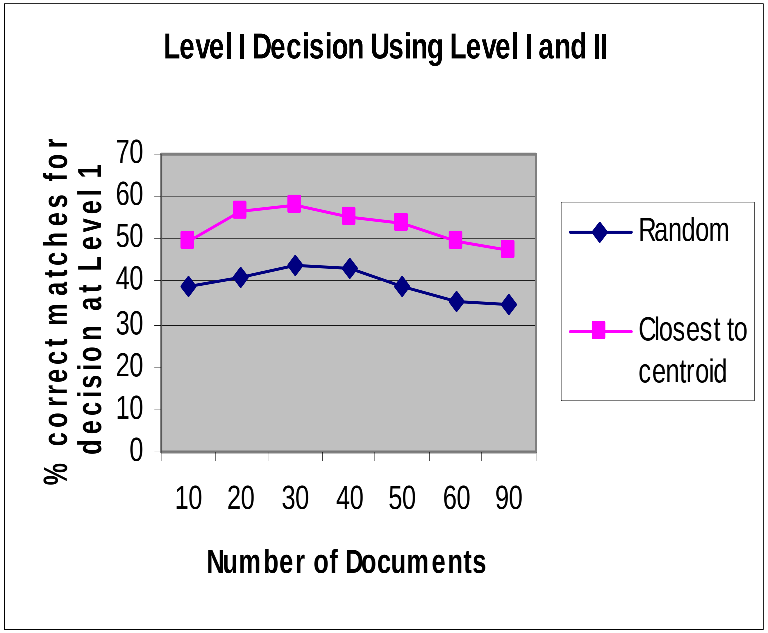
Experiment 2a – Study of Level 1 Decision

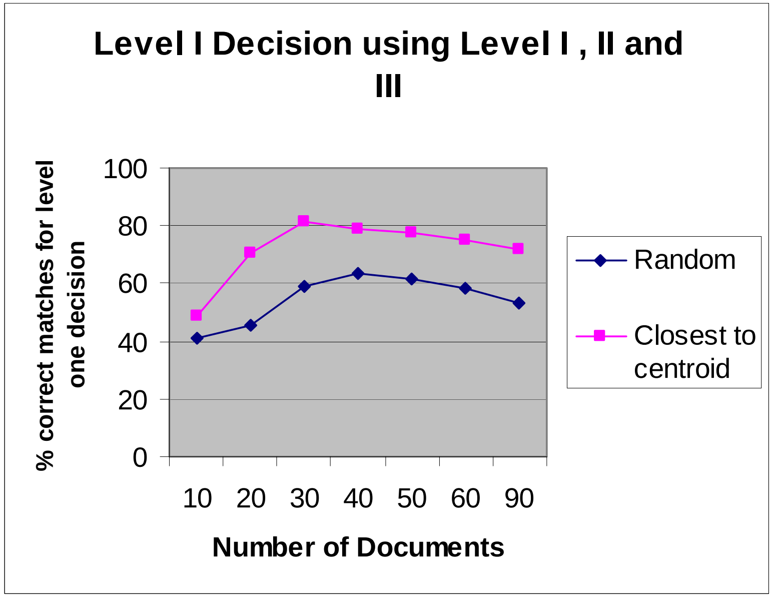
- Maximum observed accuracy 15.8% - Very Poor
- Very few documents at level-1

So, go deeper……



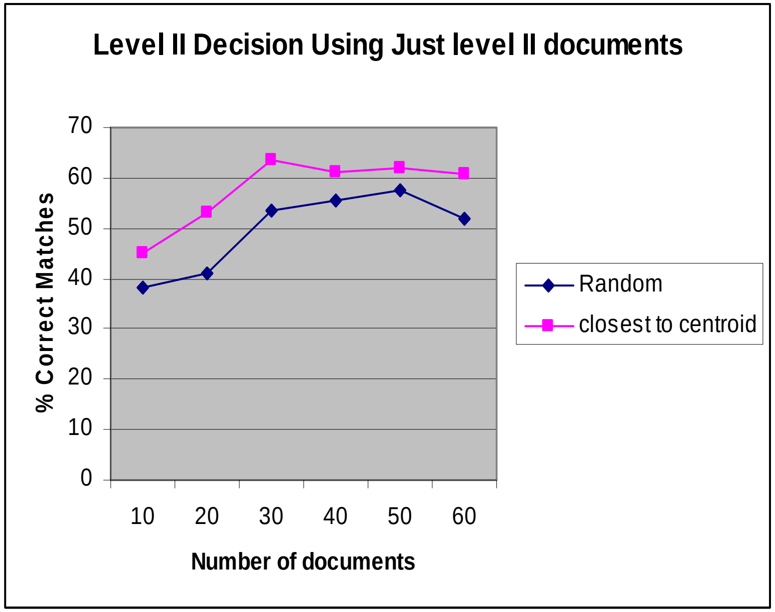
2.a: Study of Level 1 Decision.....

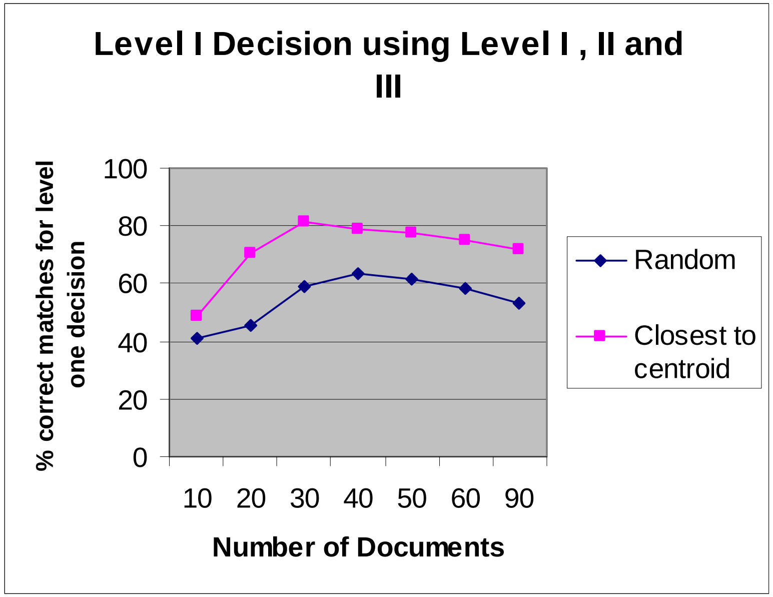




Maximum accuracy of 81.6% for level 1 decision when documents from levels 1,2 & 3 are used

Expt 2.b: Study of level 2 decision

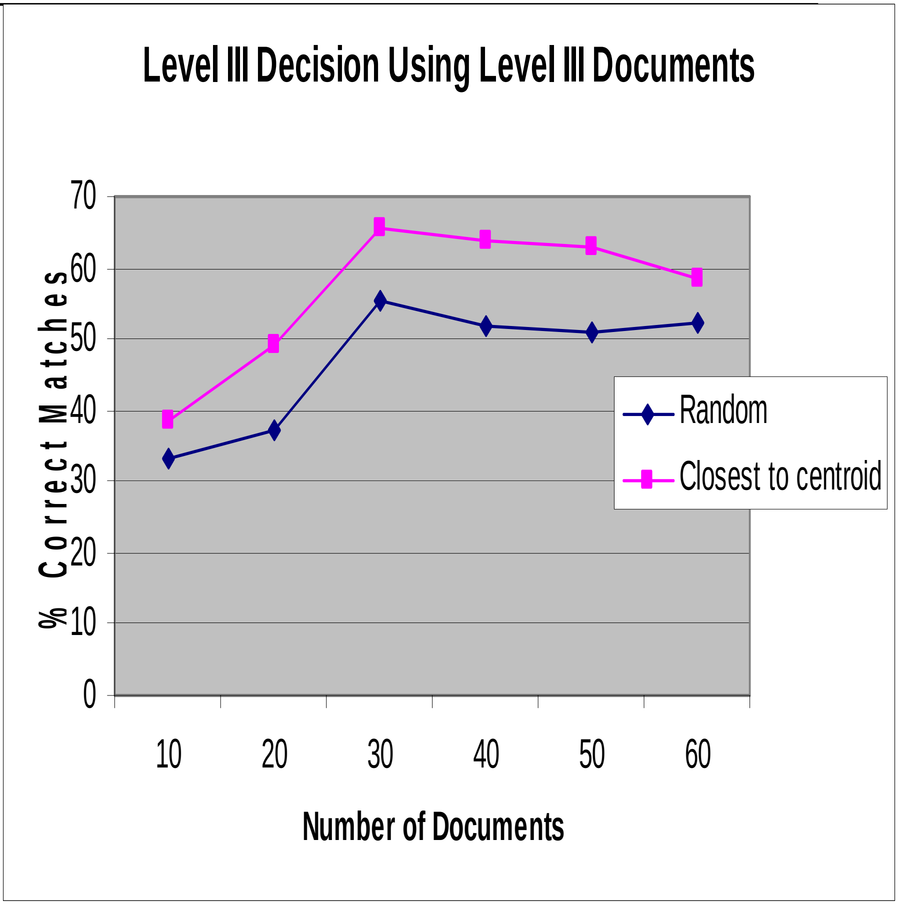




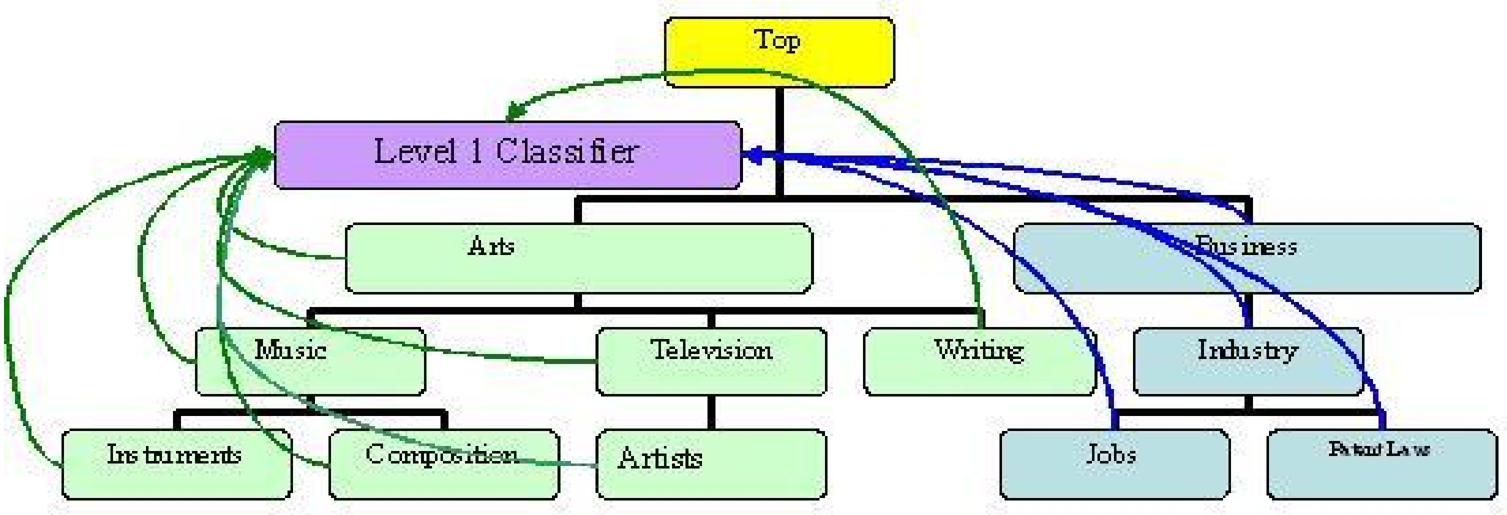
Maximum accuracy of 71.3% for level 2 decision when documents from levels 1,2 & 3 are used. 40 documents to train per concept.

Expt 2.c: Study of Level 3 Decision

- Maximum accuracy for random selection = 55.2%
- Maximum accuracy by selecting docts closest to the centroid = 65.4%
- 40.3% relative improvement over baseline

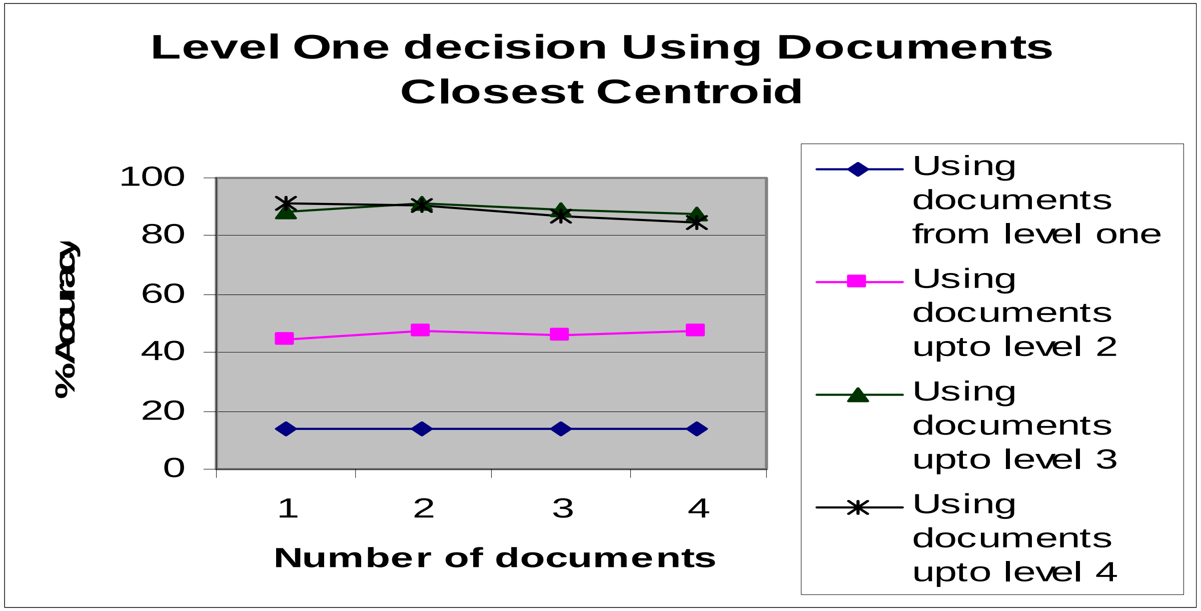


Expt 3: Effect of clustering on hierarchical classification, distributing training set across sub-concepts



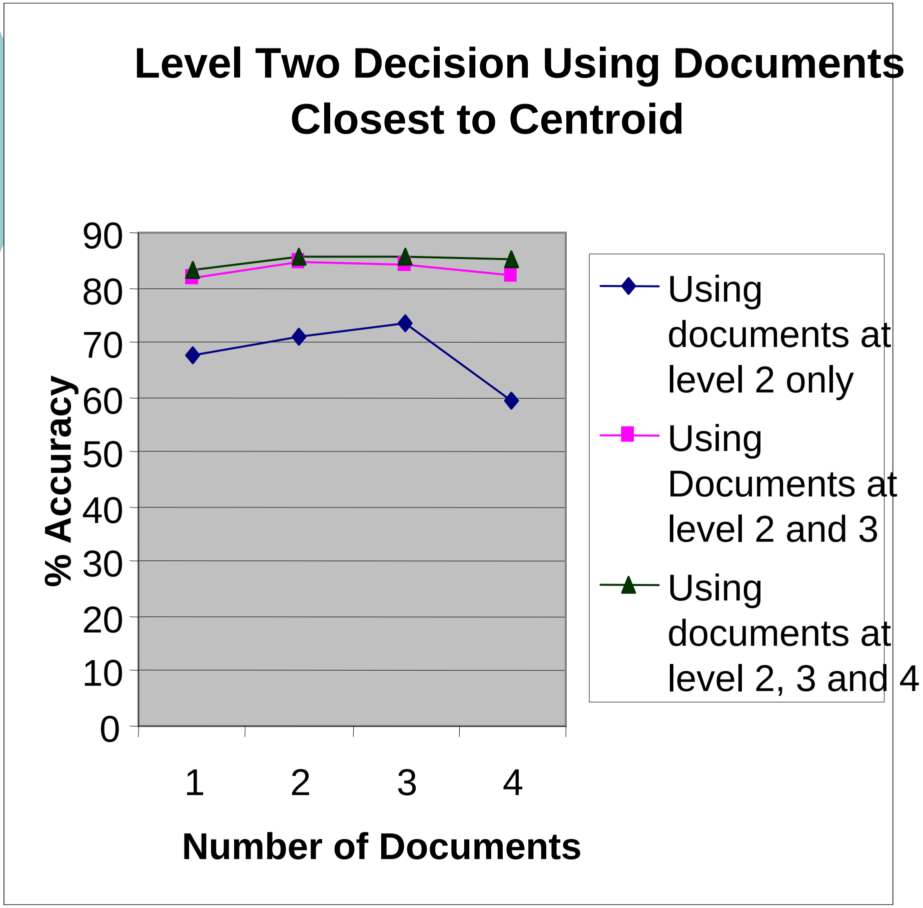
- Documents selected from each sub-concept
- Parameters we plan to tune : Depth, # of docts, random vs. closest to the centroid

Experiment 3.a: Level 1 Decision



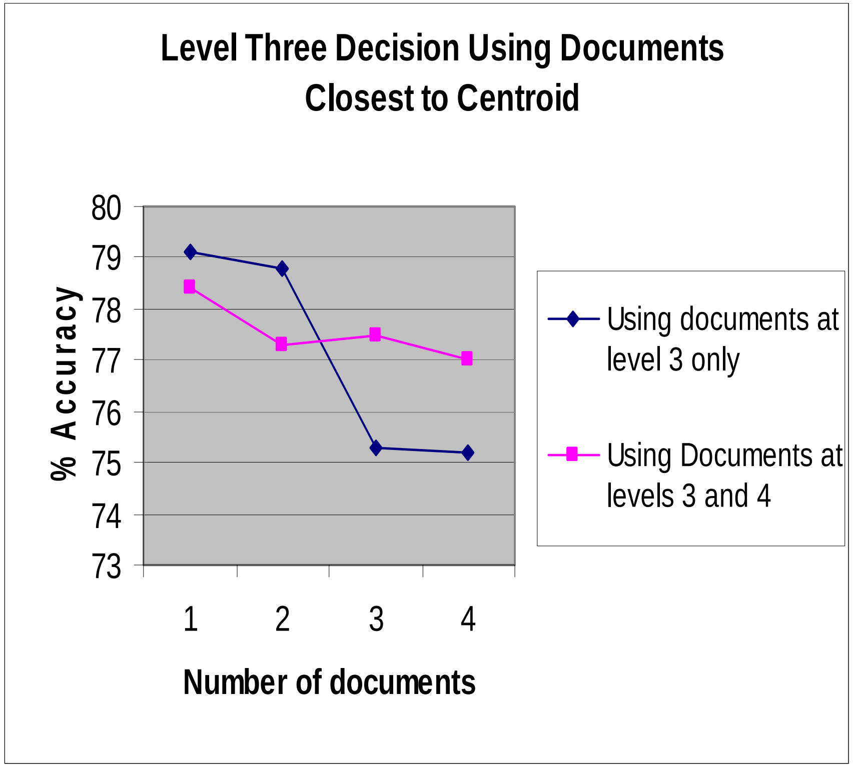
- Including level 4 – almost same results as level 3
- 91.2% Accuracy – 2 documents closest to the centroid from each concept down till level 3
- Poor results while using just level 1 or level 1 & 2

Experiment 3.b: Level 2 Decision



- ○ Using documents from levels 2&3, 2,3&4 yield almost identical results
- ○ We use till level 3 - computational time and complexity
- ○ Best observed accuracy – 84.4% - 2 docts per concept closest to the centroid

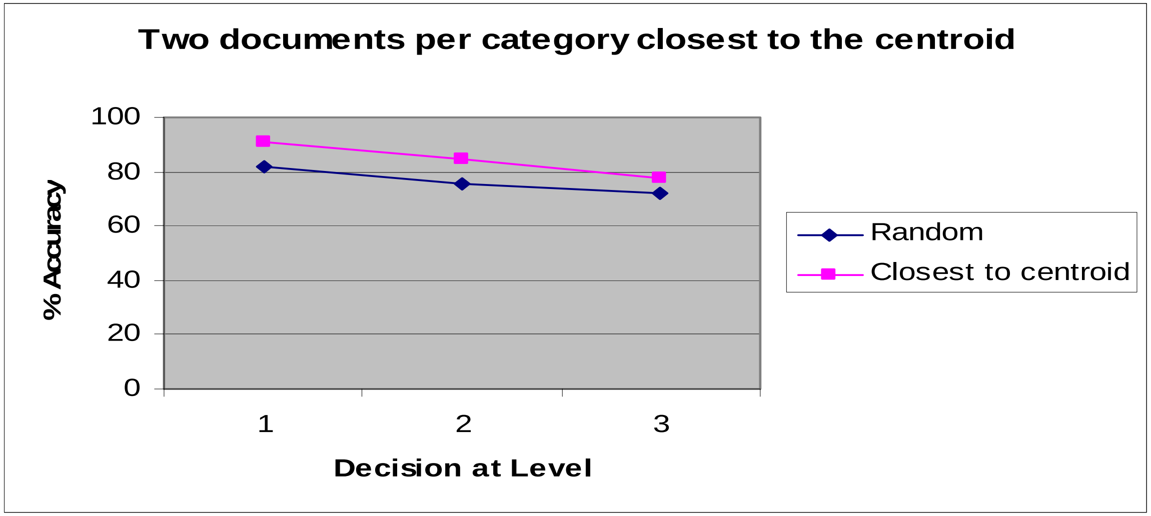
Experiment 3.c Level 3 decision



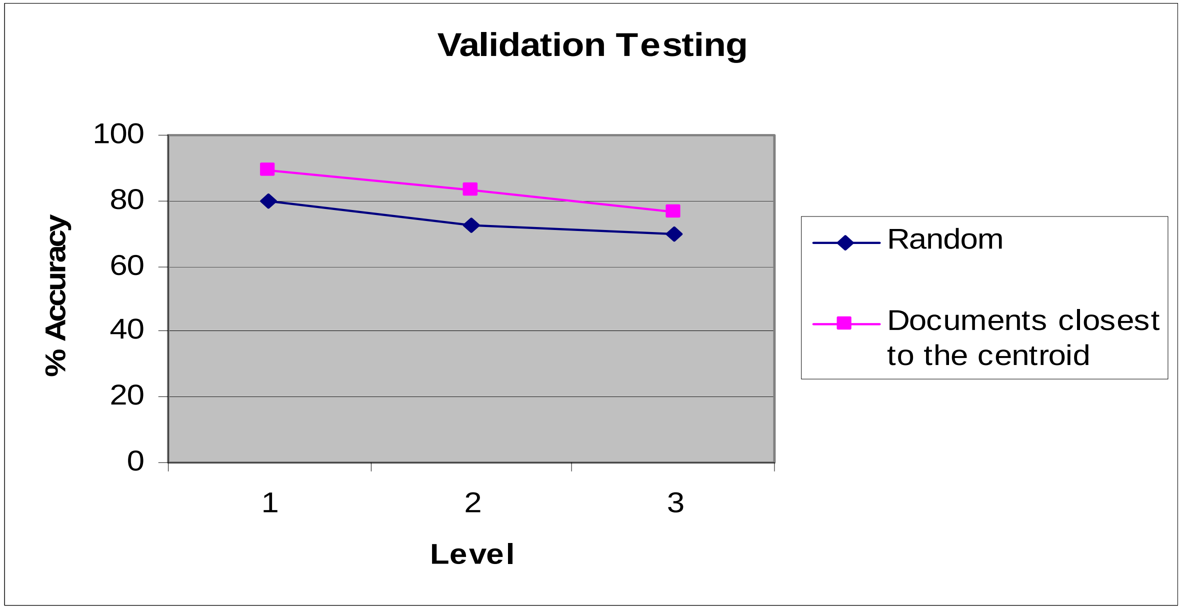
- Overall best accuracy of 79.1% at level 3 using one document from each concept that is closest to the centroid.

Training Strategy

- 2 training documents from each concept
- Down to level-3
- These documents are closest to the centroid in each concept
- Accuracy of 77.9% when we use clustering as compared to 71.8% when we select random documents

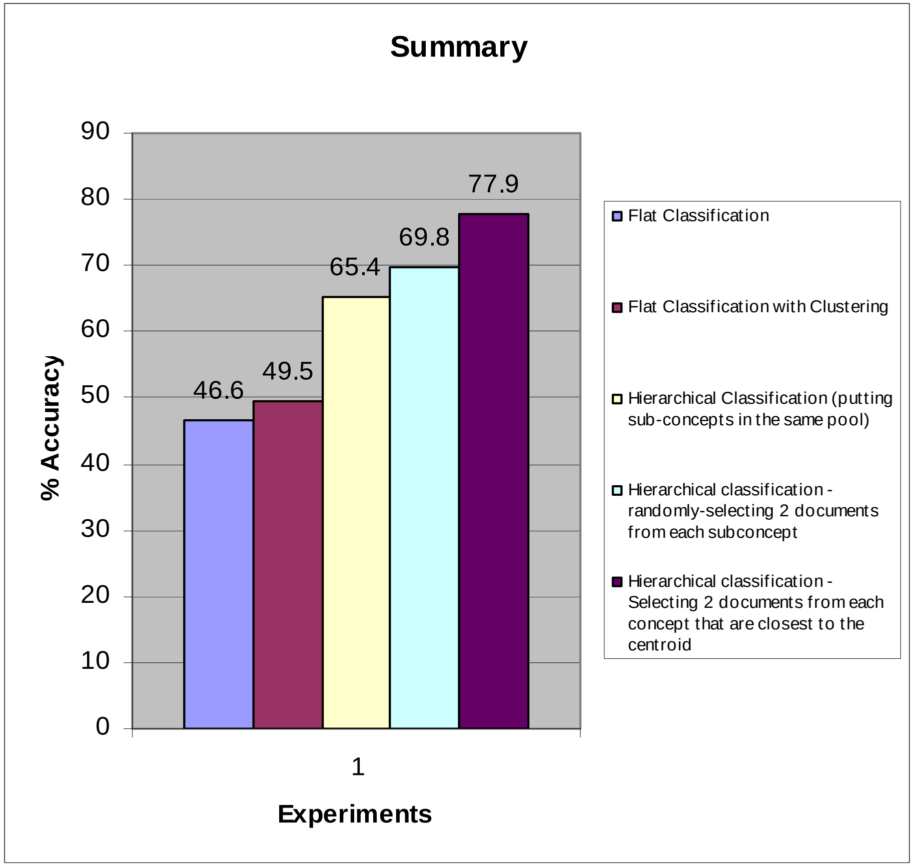


Validation Testing



- Different Test data
- Role of clustering enhances accuracy from 79.7% to 89% at level-1 and final accuracy from 69.8% to 76.2%.
- Statistically significant(*t-test value = 3.23E-05)* improvement

Conclusions



Maximum Accuracy of 77.9% when we use :

Hierarchical Classification,

2 documents closest to the centroid from each concept down till level-3 to train the classifier

Future Work

- ○ Use of other classifiers like the SVM
- ○ How to deal with the dynamic web ?
- ○ Trials on other data sets
- ○ Recovery mechanism when error is made at the parent level
- ○ Further 'divide and conquer' – Binary decisions

????’s or !!!!’s

Thank You