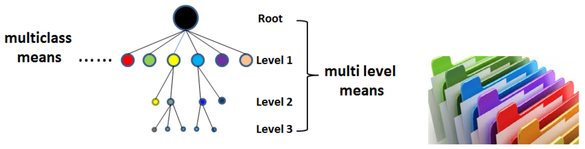
Multiclass Multi-Level Text Document Classification

Yasotha, R.[†], Charles, E.Y.A.[‡]
[†]Faculty of Graduate Studies, UoJ [‡]Department of Computer Science, UoJ

SLSWCS Student Workshop on Computer Science

Introduction

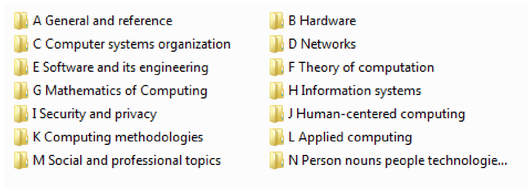


Text Document Categorization (TC) - also known as text document classification or text classification or topic spotting, is the task of assigning predefined label to text documents.

Objective

Categorize articles belong to multiclass and multi levels of topics.

- Testing platform for research study: Abstracts of conference papers displayed in ACM digital library.
- Labels: defined by ACM-CCS

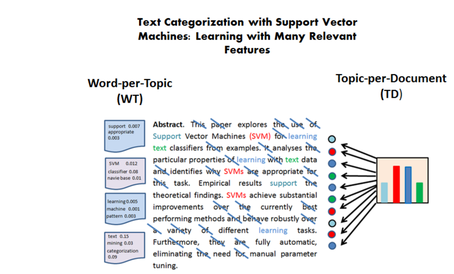


Dataset

100 distinct abstracts in each category presented in ACM-CCS in ACM digital library.

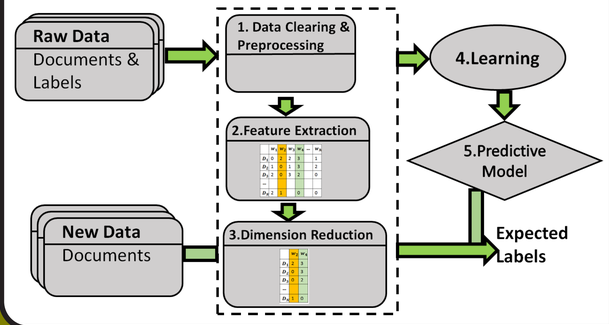
Simple Intuition:LDA Model

A document exhibits multiple topics is illustrated in the figure.



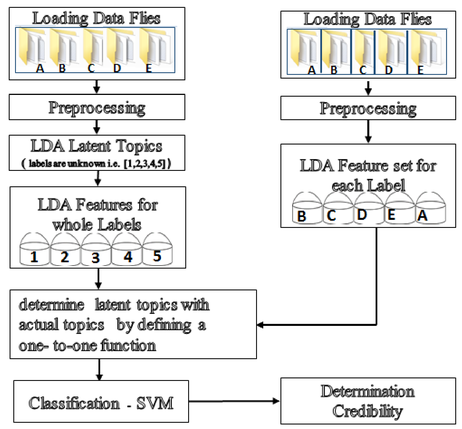
Each **topic** is a distribution over words
Each **document** is a mixture of corpus-wide topics
Each **word** is drown from one of these topics

Overview of TC process



Proposed Method

In this approach, we proposed a method to predict a label to any text documents. Here, features are extracted using topic model, LDA (Latent Dirichlet Allocation) proposed by Blei *et al.*,2003 [1,2].
In LDA, the observed data are the words of each document and the hidden variables represent the latent topical structure, ie., the topics themselves and how each document exhibits them.



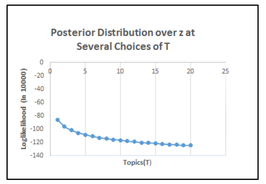
Preprocessing

1. Extract raw words.
2. Remove non-literal characters:{",<,!,',?,etc.}
3. Remove stop-words:{if, and, that,etc.}
4. Remove words that are generally used in abstracts : {paper, describe, addressing, proposed, framework, etc.}

Result

Gibb sampling algorithm is used to determine the number of clusters in the domain concerned. In this approach, samples posterior distribution at several choice of Topics (T) were computed.

The computation result is shown in the graph shows that, all level I categories under ACM-CCS are divided into four groups according to similarly of documents.



Result ...

K means algorithm is used to divide all documents under all 13 categories into four clusters and identified categories falling in each cluster.

Cluster	Categories
Cluster 0	A, H, I, J, L & M
Cluster 1	K
Cluster 2	F & G
Cluster 3	B, C, D & E

Result ...

χ^2 feature selection method is used for selecting top ten keywords from each cluster.

Cluster 0	Cluster 1	Cluster 2	Cluster 3
time	information	performance	problem
performance	data	memory	algorithm
real	user	processor	bound
data	technology	high	log
software	analysis	power	number
application	service	hardware	graph
system	process	architecture	set
language	work	level	known
algorithm	need	application	linear
specific	different	overhead	time

Result ...

The table shows the result in terms of F_1 score.

Approach	F_1 score(%)
BOW approach with NB	34.62
BOW approach with SVM	34.62
Proposed Approach	66.66

Challenges

- Size of the vocabulary
- Total number of topics
- Topics correlation: unified-topics, common-topics, similar-topics, non-overlapping-topics, etc.
- Labeling: Human factor in assigning labels to text documents in the collected dataset.

Conclusion

Instead of common practice of string matching approach, this poster presents an LDA based approach for automatic TC. Proposed model was able to categorize unseen documents with an accuracy of 66.66%.

Future Direction

- Preliminary step for automatic categorization of text documents in MLMC. This may be extended to a ontology in a poly-hierarchical structure.
- In this approach, terms are in single form was considered. Single term cannot provide complete identification of document thematic content.
- Word features in form of bigram and trigram are more influenced in the prediction of topic spotting with more effectiveness.

References

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," pp. 993–1022, 2003.
[2] D. M. Blei, "Probabilistic topic models," vol. 55, pp. 77–84, April 2012.
[3] L. G. Thoman and M. Steyvers, "Finding scientific topics," vol. 101, pp. 5228–5235, April 6 2004.