TEXT CATEGORIZATION

Amir Rahimoğlu, Çağlar Terzi,
Ömer Perçin, Emin Yiğit Köksal

Description

- Documents on the web grow rapidly
  - Over 500 billion documents
  - Infeasible to categorize them with human workforce
- Need for fast and high quality text categorization approach
- We aim to implement a document categorization algorithm based on a previous work by Chung and McLeod[1]

Motivation

- Vast amount of information and informative relations on the web
- Lots of other sites that have not yet categorized their vast amount of information
  - Ex: News sites
- Dynamic approach is required to handle high rate of data insertion
  - Web grows dynamically

Motivation

- News sites like cnn.com categorize their documents in the topic level
  - Topic: Subject; theme; a category or general area of interest
    - Ex: Airplane crash
- Also the categorization in the event level is required
  - Event: Something that happens at a given place and time a phenomenon located at a single point in space-time
    - Ex: Plane that crashed in South Korea in 1994

Example

Document	Events
1	Murder of Andrea Yates in 2001
2	Winona Ryder caught while shoplifting
3	Andrea Yates drowned her 5 children in 2001
4	Winona Ryder found guilty of shoplifting
5	Murder of Kennedy

Example

Document	Events
1	Murder of Andrea Yates in 2001
2	Winona Ryder caught while shoplifting
3	Andrea Yates drowned her 5 children in 2001
4	Winona Ryder found guilty of shoplifting
5	Murder of Kennedy

Methodology

- Pre-processing
  - Convert the words to their roots by using Porter Stemmer or another stemming algorithm
  - Remove some commonly used stop words like(have, did, and , or, etc.)

Methodology

- Categorization
  - Can also be named as clustering
  - Produce a document-term matrix
    - Data field of the matrix consists of the term weights that belong to the particular document
    - Weights of the documents are calculated as tf.idf values
  - All existing documents will be represented as a weights of the terms vector
    - By using this vector a similarity between two documents can be calculated by using cosine similarity

Methodology

- Whenever a new document is inserted to the dataset, firstly, neighbors of that particular document should be found
- Similarities between documents and the candidate document are calculated
  - If the similarity is above a certain threshold (determined experimentally)
    - That document is considered as a neighbor of the new document
- Also we will try to identify selective words of the clusters.
  - Important/unique words for that cluster

Expected Results

- Develop an efficient document categorization and suggestion system that is usable and that has high precision and recall
- Our solution should require minimal computation over existing documents when a new documents arrives (Dynamic)
- It should effectively suggest relevant documents to user when a user reads a document from the document set

References

[1] Chung, S. McLeod, D. *Dynamic Pattern Mining: An Incremental Data Clustering Approach* 2005

Thank you

- Questions?