BILKENT UNIVERSITY- IR PROJECT PLAN

TEXT CATEGORIZATION

Text Categorization, Suggestion and Selective Word Identification Approach

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

3/21/2012

1. Problem Description:

Documents on the web grow rapidly and it is infeasible to categorize them with human workforce. Generally, classical algorithms that overcome this problem either suffer from complexity or quality. Relations between documents should be identified and there is a need for fast and high quality text categorization approach. Additionally, not just static data, but also dynamic or streams of data shall be handled. In this project we aim to implement a document categorization algorithm based on a previous work by Chung and McLeod[1].

2. Motivation:

There is vast amount of information and informative relations on the web. Texts or documents are one of the most common information representation in the world.

News sites like cnn.com categorize their documents in the topic level. However there are lots of other sites that have not yet categorized their vast amount of information. Furthermore, in certain cases not just categorization in topic level, but also the categorization in the event level is required[1].

Additionally, some practical concerns must be considered. For example most of the documents in the web are uploaded dynamically. Hence a dynamic approach is required to handle high rate of data insertion. In order to handle huge amount of inserted data there is a need for an effective approach.

Moreover, usability, comprehensibility, validity, precision and recall of the categories are important factors. There should be ontology on the categories in order to create more practical categories.

In this project, we will try to propose an effective method to address these by mainly addressing the works of Chung and McLeod [1].

3. Methodology:

3.1 Pre-processing:

In this part, first we will pre-process the documents. We will convert the words to their roots by using Porter Stemmer and try to remove some commonly used stopwords like(have, did, and , or).

3.2 Categorization:

First step in the categorization algorithm is to 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 by the multiplication of the term frequency and inverse document frequency which are calculated as tf.idf values.

After creating this matrix, all existing documents are represented as a wieghts of the terms vector. By using this vector a similarity between two documents can be calculated by using cosine similarity which depends on the angle between two vectors.

Whenever a new document is inserted to the dataset, firstly, neighbours of that particular document should be found. For this purpose, similarities between documents and the candidate document are calculated, and if the similarity is above a certain threshold (exact value will be determined experimentally) then that document is considered as a neighbour of the new document, hence they will belong to the same category.

We will also try to identify selective words (important words that can signify the topic of that cluster) of the clusters.

1. Topic: Subject; theme; a category or general area of interest   e.g airplane crash
2. Event: Something that happens at a given place and time a phenomenon located at a single point in space-time.   e.g Plane that crashed in South Korea in 1994.

4. Expected Results:

We aim to develop an efficient document categorization and suggestion system that is usable and that has high precision and recall.

As we will try to implement a dynamic approach, our solution should require minimal computation over existing documents when a new documents arrives.

Also it should effectively suggest relevant documents to user when a user reads a document from the document set.

5. References:

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

2. J. Allan, J. Carbonell, G. Doddington, J. Yamron, and Y. Yang. Topic detection and tracking: pilot study final report. In *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*, 1998.

3. S. Chung, and D. McLeod. Dynamic Topic Mining from News Stream Data. In *Proceedings of the 2nd International Conference on Ontologies, Databases, and Application of Semantics for Large Scale Information Systems*, 2003.