Project Plan

A Clustering and Recommendation System for Opinion Articles in Turkish newspapers

Üstün Özgür

In this project, we aim to cluster opinion articles in Turkish newspapers and build a recommendation system on top of it. The main aim is to detect common themes in opinion articles so that a user can discover the trending topics on current day, week or month. The articles are going to be clustered around these topics, and users will be recommended articles based on their reading preferences.

Using the system, the user should be able to see on which topics most of the authors are focused today and find supporting and opposing opinions about a subject easily. For example, the system would identify that today's most important topic of discussion is the negotiations with PKK, and allow the user to see articles about that topic. The user should then be able to read all articles in the past week about this topic and possibly find new authors he might be interested in. The system should therefore be flexible in terms of time during which trends will be calculated.

This project involves a number of difficult and important steps. The first step is the scraping and content extraction from different newspapers. I currently have a system that can scrape current articles and their headings from major newspapers, I will extend this so that it can extract the main text. I could probably resort to manual text extraction for each newspaper rather than attempting a generic content extractor. Next, these raw articles should be indexed and filtered down (stemming etc) so that a fast search is possible.

Following that, two clustering techniques will be implemented for clustering the articles. Finally, a content based approach that asks the user his preferences initially and then adapts as the user likes or dislikes articles will be implemented.

The expected results are such that the system should be able to detect about 5-10 major topics within a given day. Clustering should rely on automatic means rather than being trained initially. The recommendation system might be trained as the user makes choices.

References:

Data clustering: A review, A. K. Jain, M. N. Murty, P. J. Flynn

Inverted files for text search engines, J. Zobel, A. Moffat

Web page classification: features and algorithms, X. Qi, B. D. Davison

Evaluating Collaborative Filtering Recommender Systems, Herlocker.