Instant New Event Detection and Tracking and Retrospective Event Clustering in Turkish News using Web Resources

Principal Investigator: Fazlı Can
Investigator: Seyit Koçberber
RAs supported by the project: H. Çağdaş Öcalan, Süleyman Kardaş
Budget: 124,373 YTL ($77,442 as of June 19, 2006)
Duration: July 1, 2006 – June 30, 2008

Abstract

This study aims to provide novel approaches to new event detection and tracking (NEDT) and retrospective clustering, and their prototype implementation on the Web based on multiple Turkish Web news resources. The advantages of getting news from multiple resources in an environment such as the Web can be summarized as news variety and being able to see multiple aspects of news not pieces of it. However, in multi-resource environments, news should be supplied in a meaningful context with an easy to follow presentation format. The approach presented in this project addresses the infrastructure construction of such applications. The available Web applications with similar purposes include googleNews and NewsIsFree as commercial and University of Michigan's NewsInEssence as research-oriented examples.

NEDT is a new subject in the information retrieval literature and the number of studies on Turkish in this area is next to nil. The subject made itself a place in the literature in the late 90's by following the increase in the Web provided facilities. The variety and increase in news resources due to the Web environment and what should have been done versus what has been done for Turkish and the value that would be added to the quality of daily lives explain the importance of the proposed study.

Studies to date usually try to adapt the traditional clustering algorithms to the NEDT problem. In this study for NEDT and retrospective clustering, we use the cover coefficient and the related concepts, which were developed by our contribution. Our approach to NEDT uses a new rule-based method. To use or not to use stemming and the effects of the chosen indexing approach in system effectiveness are traditional topics and continuing research problems in information retrieval. In this project, we plan to analyze the effects of different indexing techniques within the context of NEDT and retrospective clustering. The results on the effects of different approaches to system performance would shed light to future research and commercial applications. The test collection that would be built in this project would provide a standard test bed for future NEDT studies and make them easier to carry out. It would also provide an example, promote, and increase the number of such works for Turkish. The prototype Web application that we aim to develop in this study would enable us to construct and test the infrastructure of the Turkish news portal that we aim to develop in our future studies. It would also provide an example for similar studies for Turkish.

The study has qualities that can promote new research and be transferred into other application areas. Among these one can mention visualization and presentation of new events and tracked news to users, summarization of the tracked news, and applying the proposed NEDT approach to a multilingual environment. The new application areas of the approach include tracking of new developments in intelligence applications, determining new shopping trends in commercial data mining, tracking new topics in e-mail communications (e.g., in customer communications determining new problems using customer complaints), and filtering spam mails.

Keywords: cover coefficient, data mining, information retrieval, instant new event detection and tracking (NEDT), news portal, performance evaluation, retrospective clustering, rule-based decision making, Web.