Novelty Detection in Topic Tracking for Turkish and its Application to a Large-Scale News Portal
TÜBİTAK Project No. 108E074
Principal Investigator: Fazlı Can
Investigator: Seyit Koçberber
RAs supported by the project: Cem Aksoy, TBA
Budget 126,313 YTL ($106,638 as of August 12, 2008: official project approval date)
Duration: October 1, 2008 – September 30, 2010

Abstract
Multi-resource news portals provide various advantages such as richness in news, possibility of seeing news from different angles, and unbiased news presentation provided to news customers. One of the necessary important services that should be supplied by news portals is tracing and grouping of incoming stories according to their topics. The topic tracking in such systems is mainly performed in two ways: independent of users by using a new event detection and tracking (NEDT) system or by an information filtering (IF) system using news articles or words provided by users. Such news groups may contain documents with no new information due to articles containing almost the same information provided by various news providers. It is important to identify stories that include new information in topic tracking for facilitating easier topic perception by users.

In this study our aims are a) developing novelty detection (ND) methods for identifying news that contain new information in Turkish news groups generated by NEDT or IF systems, b) constructing a standard test collection for measuring the effectiveness of ND methods, c) implementing a large-scale and real-time news portal in a hardware environment that can provide a large spectrum of services with no interruption. The news portal would employ ND and other approaches that are developed in the lab environment. Observations made in the news portal environment would be provided as feedback to the research process in an iterative manner and this would provide research and application synergy. Commercial news portal examples similar to ours include GoogleNews and NewsIsFree, and research-oriented examples include Newsblaster and NewsInEssence of the Columbia University and University of Michigan, respectively. We plan to develop new approaches to ND different from the existing methods. Among these approaches, there are ones, which are based on sentence level analyses and methods based on the cover coefficient concept developed by the project proposal owners.

News web resources are frequently crawled by news portals for timely delivery of new developments. News crawling and content extraction from news pages in terms of text and pictures should be done automatically, efficiently and effectively by eliminating various noise items such as advertisements. Detection of near-duplicate documents is important since such items may affect the efficiency and effectiveness of other processes. In this project, new methods for web page content extraction and duplicate document detection will be developed and their performance will be measured and used in our large-scale news portal implementation. We plan to develop heuristic methods for news page content extraction, exploit the use of named entities in sentences for identifying near-duplicate documents.

The ND test collection that we plan to construct is like those of TDT and TREC standard test collections due to its size and coverage. It would motivate other researchers, make the research process easier, and increase the quality and quantity of studies on Turkish in this area. The planned large-scale Turkish news portal would provide effective solutions to real problems in real life environments using the scientific method. These characteristics and its new features would make it a pioneer in similar applications for Turkish.

The ND problem needs to be solved in several different fields due to web-provided or -induced information variety, information abundance, and information overlap. To the best of our knowledge, there is no ND-related study for Turkish. This fact and expected positive impact of our study by applying our research results to news portals would make a considerable effect on news consumers' quality of life, which indicates the significance of the proposal. Furthermore, the planned methods could be extended to other application areas such as summarization of tracking news; task-based information exploration in intelligence applications; detection of new developments in patient reports, e-mails and blogs; and detection of consumption trends in commercial data mining.