Bilkent News Portal: A Personalizable System with New Event Detection and Tracking Capabilities

Fazli Can, Seyit Kocberber, Ozgur Baglioglu, Suleyman Kardas, H. Cagdas Ocalan, Erkan Uyar
Bilkent Information Retrieval Group, Computer Engineering Department, Bilkent University
{canf, ozgurb, skardas, hocalan, euyar}@cs.bilkent.edu.tr, seyit@bilkent.edu.tr

Categories and Subject Descriptors

H.3.3[Information Search and Retrieval]: information filtering

General Terms

Design, Experimentation

Keywords

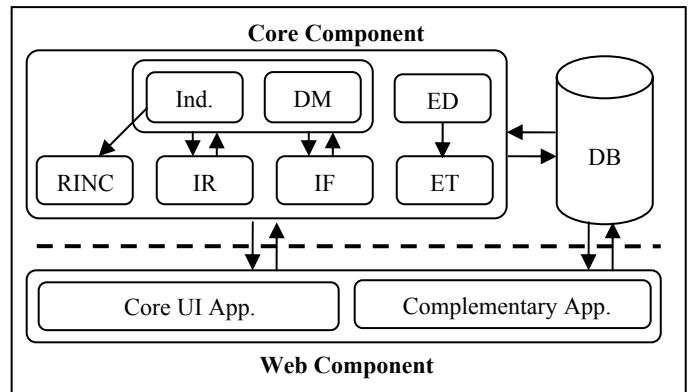New event detection and tracking, news portal, Web

1. SYSTEM OVERVIEW

Multi-source news portals, a relatively new technology, receive and gather news from several Web news providers. These systems can make the news more accessible, especially by providing event-oriented groupings by detecting and tracking the first stories of previously unseen events. In this short article we briefly demonstrate the first personalizable Turkish news portal (http://newsportal.bilkent.edu.tr/Portal) that provides the following functionalities (see Figures 1 and 2).

- New Event Detection and Tracking (NEDT): This component is based on our extensive experiments with a test collection that we constructed by downlaoding all time-stamped news articles of the year 2005 from five Web Turkish news providers. It contains more than 200,000 news and 80 events annotated by 39 native speakers. In the system implementation, for the event detection sub-component we employ the time window concept [3] and some novel approaches such as combined similarity measures.

- Information Retrieval (IR): Foundations of our IR implementation is described in [1]. In this part we extend the Lemur Toolkit (http://www.lemurproject.org/) for our purposes.

- Information Filtering (IF): Registered users are allowed to choose news that match their interests. Up to ten most recent user-selected news are employed for the generation of each IF profile using a *tf.idf* based term selection approach. Users can have several IF profiles.

- News Categorization (NC): Meta data obtained from the Web sources are used for news categorization.

- Retrospective Incremental News Clustering (RINC) : News are clustered in a restrospective and incremental manner [2]. Users can browse the cluster that contains a selected news.

- User Personalization (UP): In addition to personalized IF, users can save or send any news to the users in their friend list.

Recently we get URLs from RSS feeds of five different sources to download articles (more than 1,000 per day). In the near future, we plan to significantly increase the number of news sources and develop a task-specific crawler to download news with their pictures.

Ind.: Indexing, DM: Document Matching, ED: Event Detection,
ET: Event Tracking, UI: User Interface
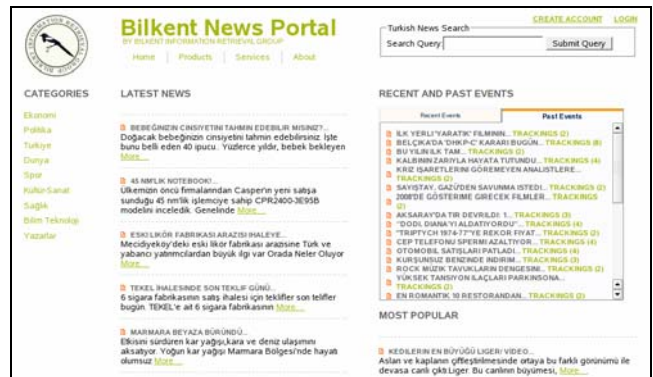**Figure 1. General system overview.**

Figure 2. Bilkent News Portal main user interface.

2. ACKNOWLEDGMENTS

3. REFERENCES

[1] Can, F., Kocberber, S., Balcik, E., Kaynak, C., Ocalan, H. C., Vursavas, O. M. Information retrieval on Turkish texts. *JASIST*, 59(3): 407-421, 2008

[2] Can, F. Incremental clustering for dynamic information processing. *ACM TOIS*, 11(2): 143-164, 1993.

[3] Luo, G., Tang, C., Yu, P. S. Resource-adaptive new event detection. *ACM SIGMOD Conf.*, pp. 497-508, 2007.